International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 504 - 508, 2013, ISSN: 2319–8656

Objects Tracking in Images Sequence Using Center-Symmetric Local Binary Pattern (CS-LBP)

Hanane. Rami
LETS/Géomat Laboratory,
Physics Department,
Mohamed V University,
Faculty of Science,
Rabat, Morocco

Mohammed. Hamri
LETS/Géomat Laboratory,
Physics Department,
Mohamed V University,
Faculty of Science,
Rabat, Morocco

Lhoucine. Masmoudi
LETS/Géomat Laboratory,
Physics Department,
Mohamed V University,
Faculty of Science,
Rabat, Morocco

Abstract: In this paper we present a method for objects tracking in images sequence. This approach is achieved into two main steps. In the first one, we constructed the Center-Symmetric Local Binary Pattern (CS-LBP) histogram pattern of each image in the sequence and the reference pattern. In the second one, we perform the algorithm by the pattern selected based on a distance measures to find similarity between two histograms. The maximum CS-LBP histogram distance gives best results than the chi-square one. The proposed approach has been tested on synthetic and real sequence images and the results are satisfactory.

Keywords: Sequence image, Computer vision, Tracking, LBP, CS-LBP histogram, Chi-square distance..

1. INTRODUCTION

Tracking systems is important in computer vision. It is applied in different domain, for example in video surveillance and human computer interfaces (HCI). Various methods can be found in the literature and can be roughly classified into two basic categories:

- The first category is the algorithms that estimate the absolute positions of the pixel in each image independently. This category includes the center-of-mass, or centroid algorithm [1,2] and direct fits of Gaussian curves to the intensity profile [3,4].

- The second category includes algorithms that estimate the change in position of a pixel by comparing an image to one subsequent. This category includes cross-correlation method [5, 6, 7], and SAD algorithm method [8]. The use of the second category is well known and commonly used in tracking for sequence image. And it is commonly used in tracking vision for the visual matching problem [9].

- The SAD method suffers from the sensitivity to intensity scaling of the image and the template [11, 12, 13], but the ZNCC method presents an ambiguity in the area with similar brightness or similar texture. It is demonstrated that the CS-LBP is mainly characterized by the invariance to monotonic changes in gray-scale and fast computation, and it has proven performance background in texture Classification [10]. While operating in gray-scale color space, CS-LBP is also robust to illumination changes.

- Texture, which has not enjoyed major attention in tracking applications, provides a good option to enhance the power of color descriptors. In this way we propose to use the CS-LBP [31] histograms to tracking the motif in a sequence of images. In order to show the feasibility of the proposed method, it is tested and applied to both real image sequences and synthesized image sequences

2. LOCAL BINARY PATTERNS (LBP)

The basic local binary pattern operator, introduced by Ojala et al. [15,16], was based on the assumption that texture has locally two complementary aspects, a pattern and its strength. In that work, the LBP was proposed as a two-level version of the texture unit [17,14] to describe the local textural patterns. The original version of the local binary pattern operator works in a 3 × 3 pixel block of an image. The pixels in this block are thresholded by its center pixel value, multiplied by powers of two and then summed to obtain a label for the center pixel.

As the neighborhood consists of 8 pixels, a total of 2^8 = 256 different labels can be obtained depending on the relative gray values of the center and the pixels in the neighborhood. See Fig.1 for an illustration of the basic LBP operator.

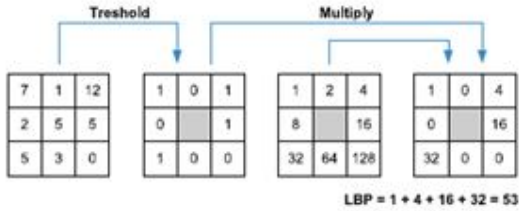


Fig. 1: The original LBP

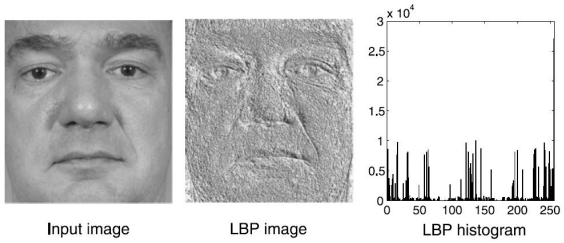


Fig.2 : Example of an input image, the corresponding LBP image and histogram

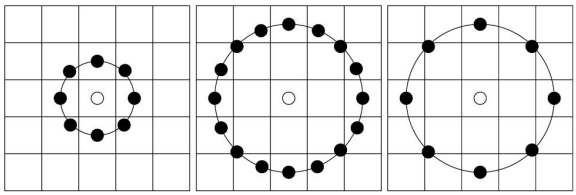


Fig.3 : The circular *(*8, 1*)*, *(*16, 2*)* and *(*8, 2*)* neighborhoods. The pixel values are bilinearly interpolated whenever the sampling point is not in the center of a pixel

Local binary pattern is a simple description operator of local texture. It can resist to the changes of illumination [10, 11]. And it has proven performance background in texture classification [10]. In recent years, the LBP operator has been used for texture classification, face recognition, image retrieval and other fields.

2.1 Center-Symmetric LBP

Center-Symmetric Local Binary Patterns (CS-LBP) [31] were developed for interest region description. CS-LBP aims for smaller number of LBP labels to produce shorter histograms that are better suited to be used in region descriptors. Also, CS-LBP was designed to have higher stability in flat image regions.

In CS-LBP, pixel values are not compared to the center pixel but to the opposing pixel symmetrically with respect to the center pixel. See Fig. 4 for an illustration with eight neighbors [31].

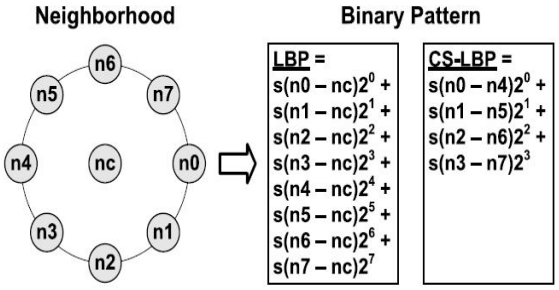


Fig 4 : LBP and CS-LBP features for a neighborhood of 8 pixels

3. HISTOGRAM DISTANCE

The use bin-to-bin distances for comparing histograms is very important. This practice assumes that the histogram domains are aligned. This distance depends on the number of bins. If it is low, the distance is robust, but not discriminative, if it is high, the distance is discriminative, but not robust. Distances that take into account cross-bin relationships (cross-bin distances) can be both robust and discriminative. There are two kinds of cross-bin distances. The first is the Quadratic-Form distance [19]. Let P and Q be two histograms and A the bin-similarity matrix. The Quadratic- Form distance is defined as:

$$QF^{A}(P,Q) = \sqrt{(P-Q)^{T} A(P-Q)}$$

When the bin-similarity matrix A is the inverse of the covariance matrix, the Quadratic-Form distance is called the Mahalanobis distance. The second type of distance that takes into account cross-bin relationships is the Earth Mover's Distance (EMD) [20].

In many natural histograms the difference between large bins is less important than the difference between small bins and should be reduced. The Chi-Squared (***X2***) is a histogram distance that takes this into account. It is defined as:

$$\chi^{2}(P,Q)=\frac{1}{2}\sum_{i}\frac{(P_i-Q_i)^2}{(P_i+Q_i)}$$

The ***X2*** histogram distance comes from the x2 test-statistic [21] where it is used to test the fit between a distribution and observed frequencies. Chi-square histogram distance is one of the distance measures that can be used to find dissimilarity between two histograms. ***X2*** was successfully used for texture and object categories classification [22, 23, 24], near duplicate image identification [25], local descriptors matching [26], shape classification [27, 28] and boundary detection [29].

4. PROPOSED METHOD

The proposed method is achieved in two main steps. In the first one, we constructed the Center-Symmetric Local Binary Patterns (CS-LBP) [31] histogram pattern of each image in the sequence and the reference pattern. In the second one, we perform the algorithm by the pattern selected based on a distance measures to find similarity between two histograms. The flowing we present the algorithm used in this stud.

Algorithm:

First step

1. Extract reference pattern

2. Calcul CS-LBP histogram of reference pattern:
   H(i,j) = HistogramPattern(i,j)

3. Extract pattern of each image in the sequence

4. Calcul CS-LBP histogram pattern of each image in the sequence :
   Hn(i,j)= Histogram NewImage(i,j)

End

Second step

Medd=Abs(max(H(i,j))- max(Hn(i,j)))

Min(Medd(I,j))

Or

Calcul the Chi-square distance between H(i,j) and Hn(i,j)
Min(chi2(H(i,j), Hn(i,j))

End

5. EXPERIMENTAL RESULT

In this section, we present the experimentation results of into tracking image sequences. Real image sequences and synthesized image sequences are considered. For evaluation the algorithm tracking results we use Euclidean distance.

5.1 Synthesized Sequence Image

We used a sequence of grayscale image containing a moving ball. This database gives returns to Strauss [18]. Table 1 shows the cumulative Euclidean distance r from the pixel position for each images i and (i +1) for the similarity measure using chi-square and maximum histogram.

Table 1: Cumulative Euclidean distance for the two method

Sequence N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Chi-square	4	77	122	158	168	236	273	299	316	369	386	403	443	479	508	545	571	611	660
Max-histogram	1	54	91	100	101	166	192	205	250	299	316	336	373	409	435	472	512	537	586

Figure 5 presents the evolution of the position of pixels for each image using de cumulative Euclidean distance. The figure shows that the similarity measure uses the maximum histogram is below the chi-square. Therefore, the maximum CS-LBP histogram distance gives best results than the chi-square one. Finally, we present examples of image sequence using the max histogram. From the results, it can be seen that its performance is acceptable for the synthetic sequence images.

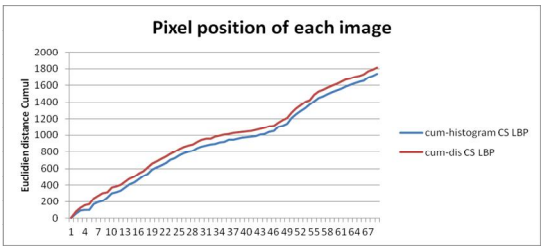


Figure 5 :Evolution of the position of pixels for each image using de cumulative Euclidean distance

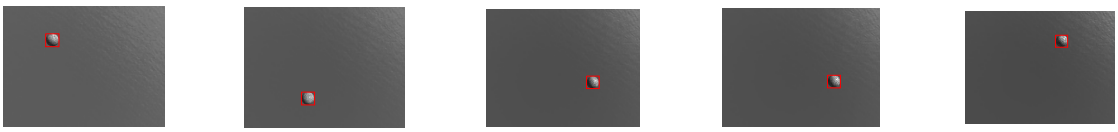


Image1 Image15 Image30 Image45 Image 60

Fig. 6 : same image using max-histogram for tracking the ball

5.2 Real Sequence Image

In order to compare the performances of the method we have considered a real sequence image. We have used the video realized by Sargi [30] for tracking the face. Table 2 present the cumulative Euclidean distance from the pixel position for each images i and (i +1) for the similarity measure using chi-square and maximum histogram. The table shows that the values relate to the measurement of Chi-square augment rapidly. In particular; the both methods are almost similar in images 4 and 5 as you can see in Fig.7.

Table 2: Cumulative Euclidean distance for the two method

Image Sequence N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Chi-square	0	100	100	105	105	157	207	211	211	279	352	501	502	502	502	520
Max-histogram	0	26	26	99	99	119	119	120	120	188	273	318	386	402	402	442

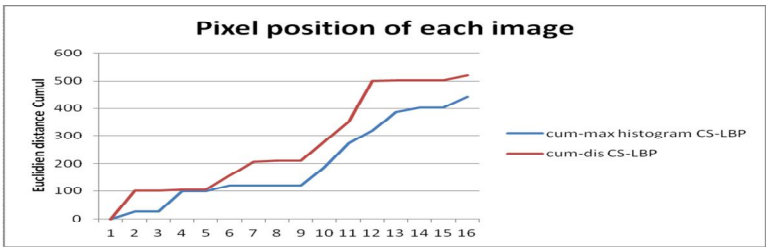


Fig. 7 :Evolution of the position of pixels for each image using de cumulative Euclidean distance

Image1 Image5 Image 10 Image15 Image 17

Fig. 8 : same image using max-histogram for tracking the face

6. CONCLUSION

In this paper a method for objects tracking in images sequence using Center-Symmetric Local Binary Patterns (CS-LBP). For evaluation the algorithm tracking results we use the cumulative Euclidean distance from the pixel position for each images. The maximum CS-LBP histogram distance gives best results than the chi-square one. From the results, it can be seen that its performance is acceptable for both the synthetic and real sequence images. In the future work, we will exploit the information color for constructed the CS-LBP operator.

7. REFERENCES

[1] Ghosh and Webb. 1994. Automated detection and tracking of individual and clustered cell surface low density lipoprotein receptor molecules.Biophys. J.66:1301–1318.

[2] Lee, and al. 1991. Direct observation of Brownian motion of lipids in a membrane.Proc. Nat. Acad. Sci. U.S.A.88:6274 – 6278.

[3] Anderson, and al. 1992. Tracking of cell surface desceptors by fluorescence digital imaging microscopy using a charge-coupled device camera.J. Cell ci.101:415– 425.Tavel, P. 2007 Modeling and Simulation Design.
AK Peters Ltd.

[4] Smith, and al. 1999. A direct comparison of selectin-mediated transient, adhesive events using high temporal resolution.Biophys. J.77:3371–3383.

[5] Gelles, and al.1988. Tracking kinesin-driven movements with nanometre-scale precision.Nature.331:450 – 453.

[6] Kusumi, and al. 1993. Confined later diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). Effects of calcium-induced differentiation in cultured endothelial cells. Biophys. J.65:2021–2040.

[7] Guilford and Gore, 1995. The mechanics of arteriole interstitium interaction. Microvas. Res.50:260 –287.

[8] Vanne, and al, 2006. "A High-Performance Sum of Absolute Difference Implementation for Motion Estimation," Circuits and Systems for Video Technology, IEEE Transactions on, vol.16, no.7, pp.876-883.

[9] Kanade and Okutomi.1994. A stereo matching algorithm with an adaptive window: theory and experiment. IEEE Transactions for Pattern Analysis and Machine Intelligence 16, 920-932.

[10] Ojala and al, 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7): 971-987

[11] Cho, and Yun, 2005 .Selective-attention correlation measure for precision video tracking. IEICE Trans. Inf. Syst. E88-D(5), 1041–1049

[12] Bohs, and al.1999. Speckle tracking for multi-dimensional flow estimation. Ultrasonics 38, 369–375 (2000)

[13] Heikkil and Pietik ,2006. A texture-based method for modeling the background and detecting moving objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28:657–662

[14] Timo Ahonen and al, 2009. Rotation invariant image description with local binary pattern histogram

[15] Ojala T, Pietikäinen M & Harwood D (1996) A comparative study of texture measures with classification based on featured distribution. Pattern Recognition, 29(1):51-59.

[16] A. Hadid, M. Pietikainen and T. Ahonen. A Discriminative Feature Space for Detecting and Recognizing Faces. Proc of CVPR 2004.

[17] Jo Chang-yeon, "Face Detection using LBP features," Final Project Report, December 2008.

[18] Olivier STRAUSS Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier Département Robotique LIRMM 161, Rue ADA 34392 Montpellier CEDEX 5 France

[19] Hafner, J., Sawhney, H., Equitz, W., Flickner, M., Niblack, W.: Efficient color histogram indexing for quadratic form distance functions. PAMI (1995)

[20] Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. IJCV (2000)

[21] Snedecor, G., Cochran, W.: Statistical Methods, ed 6. Ames, Iowa (1967)

[22] Cula, O., Dana, K.: 3D texture recognition using bidirectional feature histograms. IJCV (2004)

[23] Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV (2007) 3

[24] Varma,M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. PAMI (2009) 3

[25] Xu, D., Cham, T., Yan, S., Duan, L., Chang, S.: Near Duplicate Identification with Spatially Aligned Pyramid Matching. CSVT (accepted) 3

[26] Forss´en, P., Lowe, D.: Shape Descriptors for Maximally Stable Extremal Regions. In: ICCV.(2007) 3

[27] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI (2002) 3, 11

[28] Ling, H., Jacobs, D.: Shape classification using the inner-distance. PAMI (2007) 3, 11

[29] Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI (2004).

[30] Mehmet Emre Sargin and al.2005. Combined Gesture-Speech Analysis and Synthesis eNTERFACE05 Workshop in Mons, Belgium.

[31] Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recognit. 42(3), 425–436 (2009)

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 509 - 516, 2013

Development of GRID Portal

B. Dhana Lakshmi
Department of Computer Science and Engineering
ARYABHATA Institute of Technology & Science
Jawaharlal Nehru Technological University
On Srisailam Highway, R R Dist, Hyderabad, India

Abstract**:** The objective of this paper is to develop a framework for submitting the jobs and integrates those jobs to available systems by using a meta-scheduler. The meta-scheduler being used is the GridBus broker and the middleware being used is Globus. The GridSphere Portal is used to submit the jobs to the GridBus broker which in turn submits jobs to systems running Globus middleware by using scheduling algorithms and results are retrieved from them. Grid may consist of normal and scheduled PCs and it can also increase the number of systems.

Keywords*:* *g*rid; meta scheduler; resource broker; middleware; grid sphere.

1. INTRODUCTION

A grid [1] is a collection of distributed resources that facilitates the sharing, distribution and aggregation of services depending on their performance and quality-of-service measurements. A resource on a grid could be any entity that provides access to a service. This could range from servers to databases, scientific instruments, applications and the like.

Users make use of resource brokers [2] to know the transparency of heterogeneous resources in the field of Grids and distributed systems. Meta-scheduler is one that is specially designed to schedule jobs across various grid middleware. The various grid middleware accept these jobs from the meta scheduler, process them, and return the results back to the meta-scheduler.

A meta-scheduler schedules work across a number of clusters or grids, each of which has its own independent scheduling solutions. It accepts the different application level local schedulers and employs the scheduling policies between different applications. Thus, this paper will be used to submit and execute the jobs from different applications by giving the results to applications.

2. PROPOSED SCHEME

In this scheme, the main objective is to integrate the GridSphere Portal with grid middleware's by using a meta scheduler. User submits his job through the GridSphere Portal which is submitted to the meta-scheduler.

The meta scheduler that we can use is GridBus Broker which is responsible for submitting the job to the middleware. Globus will act as a middleware. Once the jobs are run on the middleware the results are returned to the meta-scheduler which in turn returns to the portal and then to users to check the output.

The user can submit the job to the middleware of his/her choice and he provides the path of the application file which has to be run on the middleware and he can also specify the middleware on which he wants his job to be running. The application file is the actual application to be run and the resource file consists of all the resources available and the credentials required for accessing these resources.

3. GRID SPHERE PORTAL

It is developed using the GridSphere Portal Framework, which consists architecture for deploying "pluggable" web applications using the Portlet model that provides development environment for easily creating new portlet applications and offers a core set of portlets for the management of portlets, users, groups and layouts.

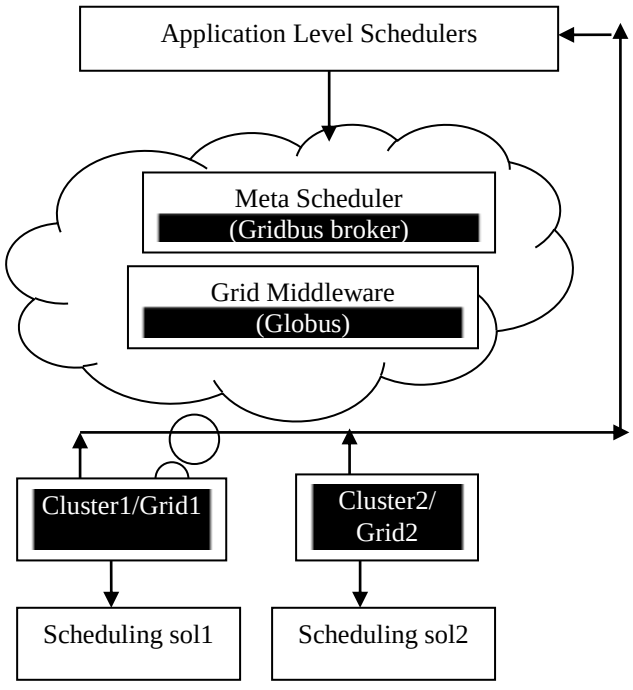


Figure 1. Gridsphere Portal

4. SYSTEM DESIGN

System design is implemented by using the following components.

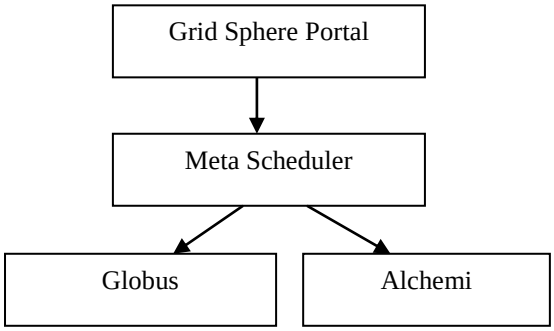


Figure 2. Proposed Design Architecture

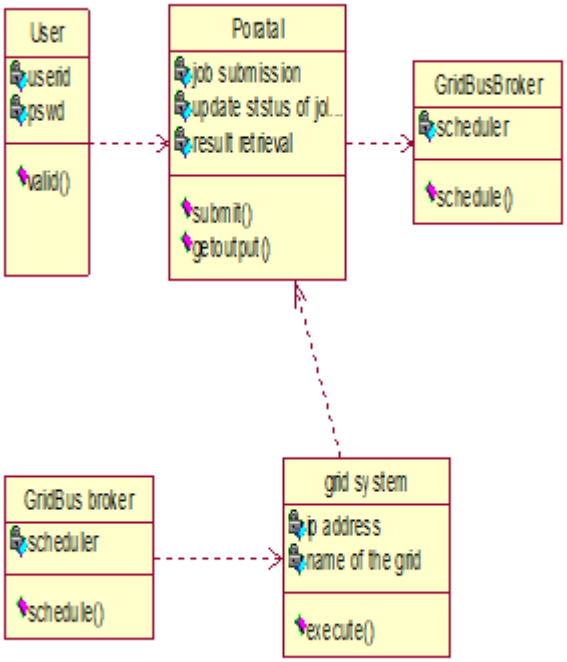


Figure 3. Gridsphere Portal Class Diagram

4.1 Gridbus Broker

It is used to support [2] both computational and data grid applications and its architecture have importance on simplicity, and extensibility in executing the grid applications. Transparency is maintained in running grid nodes. Set the proxy certificates before running the broker and the device groups have to be manually targeted before running the broker as the jobs will not be dispatched otherwise.

When running the broker, it is needed to input the resource description, which specifies the available resources and describes their attributes. The broker is pointed to the location of the resource description file which contains the description of the resources that are to be used by the broker for executing the grid applications.

This file may be modified to specify the resources the user has access to. Each run of the broker creates a separate dimension.

4.2 Broker Properties Configuration

The broker is configured in two ways, either by providing it configuration values or by creating an object to broker. Resources are classified into three types as compute, storage and service resources. Compute resources are servers to which the user's jobs are submitted for execution.

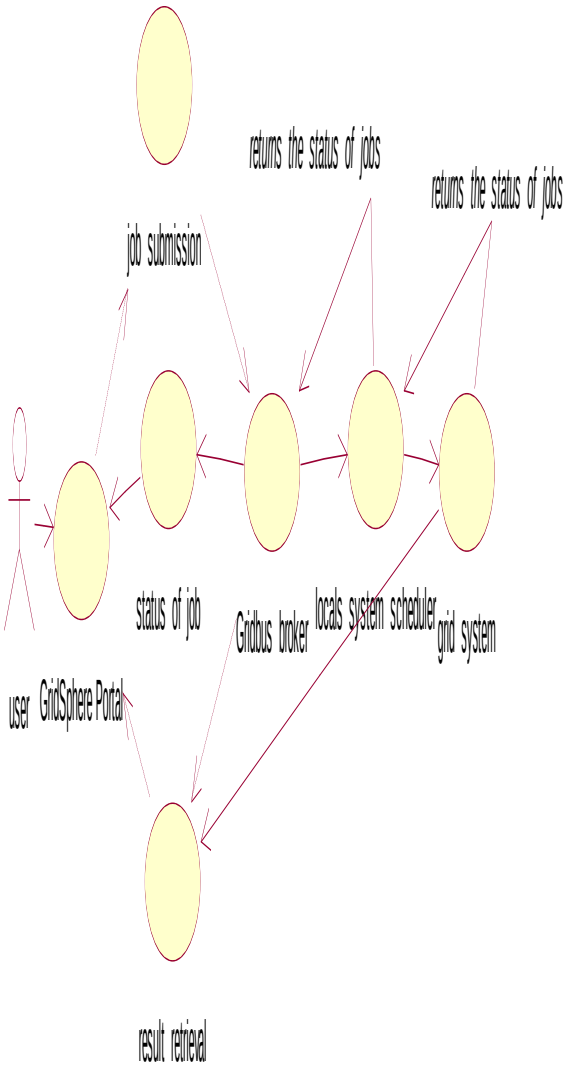


Figure 4. Gridsphere Portal Usecase Diagram

Storage resources are used to store the results of execution and hence can be considered as data sinks and which provide generic services that can be used by the broker. A service resource can be of two types-information services and application services.

Application services provide applications hosted on nodes that can be accessed as a service. Information services are entities which provide information about other resources or services.

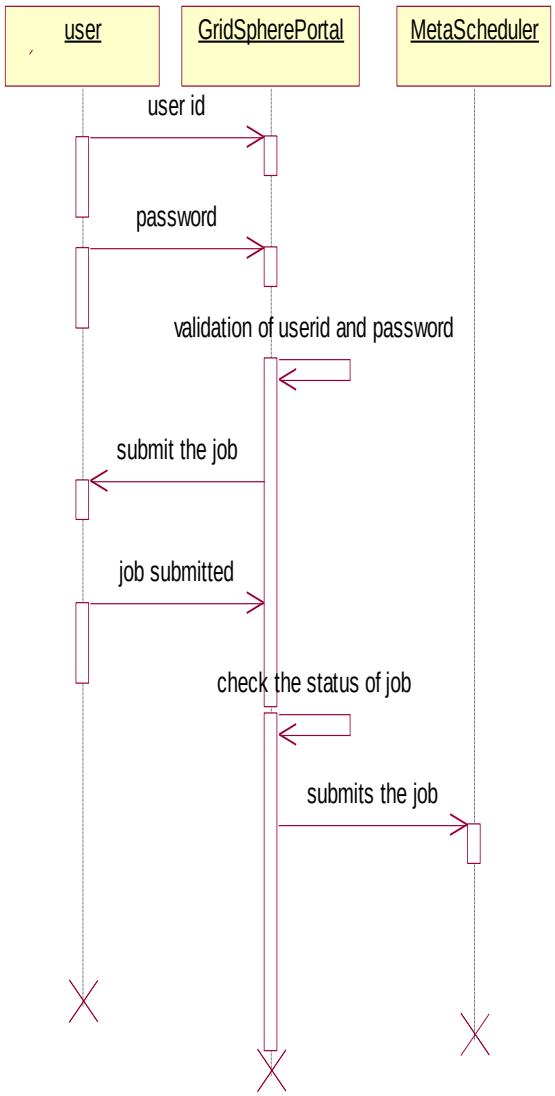


Figure 5. Gridsphere Portal Sequence Diagram1

A compute resource is associated with a domain that can take two values local and remote. Local resources could be the local computer or a cluster, on which the broker runs. Remote compute resources are used to represent nodes on the grid which have a job submission interface accessible via a network.

4.3 Broker Properties Configuration

The broker is configured in two ways, either by providing it configuration values or by creating an object to broker. Resources are classified into three types as compute, storage and service resources. Compute resources are servers to which the user's jobs are submitted for execution.

Storage resources are used to store the results of execution and hence can be considered as data sinks and which provide generic services that can be used by the broker. A service resource can be of two types-information services and application services.

Application services provide applications hosted on nodes that can be accessed as a service. Information services are entities which provide information about other resources or services. A compute resource is associated with a domain that can take two values local and remote. Local resources could be the local computer or a cluster, on which the broker runs. Remote compute resources are used to represent nodes on the grid which have a job submission interface accessible via a network.

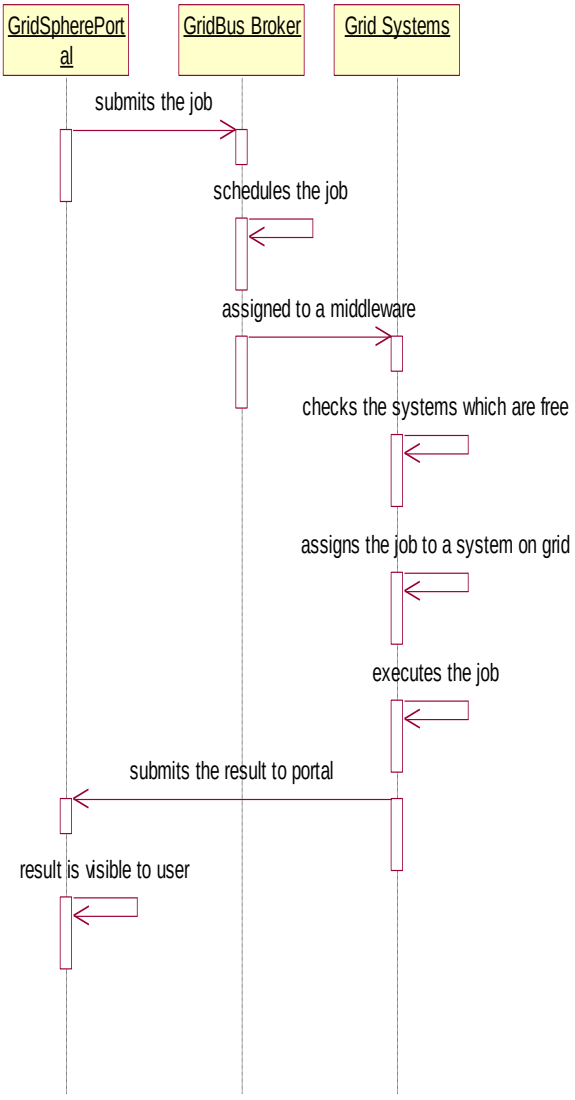


Figure 6. Gridsphere Portal Sequence Diagram2

4.4 Resource Description

This is needed by the Broker, which is used to describe two types of entities resources and to access the resources.

4.5 Broker Entities

4.5.1 Farming Engine

It is the central component which maintains the overall state of the broker at all times. It is the glue that binds all the components together. It acts as a container for the job and server collections.

It is the component that interacts with external applications and initiates the scheduling. The farming engine can be considered as the broker's in-memory database, as it holds the broker's current state at any point of time.

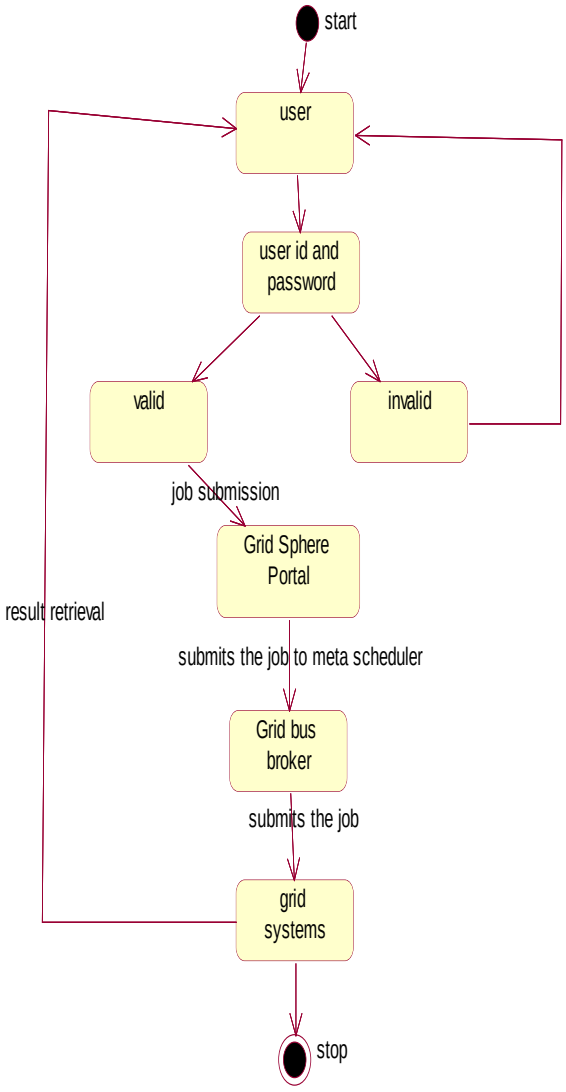


Figure 7. Gridsphere Portal Activity Diagram

4.5.2 Scheduler

Mapping of jobs to nodes in a grid platform is done by an autonomous component [1] which is known as a scheduler. It can schedule jobs based on metrics which do not depend on the underlying platform.

4.5.3 Job

It is an abstraction for a unit of work assigned to a node. It consists of variables and a task. A variable [1] consisting of a single value for a job, is called a single variable. A task is the description of what has to be done by the job.

It is composed of a set of commands. Job monitor is responsible for monitoring the execution of all jobs submitted to the node corresponding to the middleware.

4.5.4 Data Hosts & Data File Objects

These are nodes on which data files have been stored. These objects store the details of the data files that are stored on them such as their path on the disk and the protocol used to access them.

The data host objects also maintain a list of the compute resources sorted in the descending order of available bandwidth from the host. Data file objects store attributes of input files that are required for an application such as size and location. A data file object links to the different data hosts that store the file. Overall, the broker is designed to be a loosely coupled set of components working together.

4.5.5 Algorithm Steps

Algorithm for submission and monitoring cycle for job sent to a Globus node consists of following steps..

1. The scheduler submits a job to a server.
2. The server puts the job in its local job buffer and informs to server manager.
3. The server manager creates a middleware specific agent and sends it to remote machine.

4.5.6 Reaalization as a Framework.

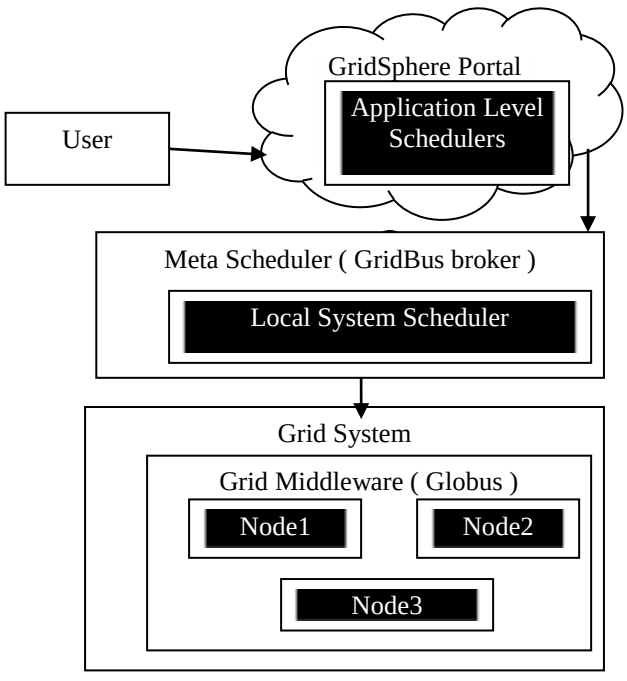


Figure 8. Basic Design Model

5. RESULTS & DISCUSSION

5.1 Resource Portlet

This portlet is used to select the resources on which the user wants to submit his job. The resources are identified by the name and type of resource.

Figure 9. Resource Portlet

5.2 Application Portlet

This is a portlet from where the user submits his job. User provides the job as a file and the application to be run. The job, which he submits, is sent to the GridBus Broker.

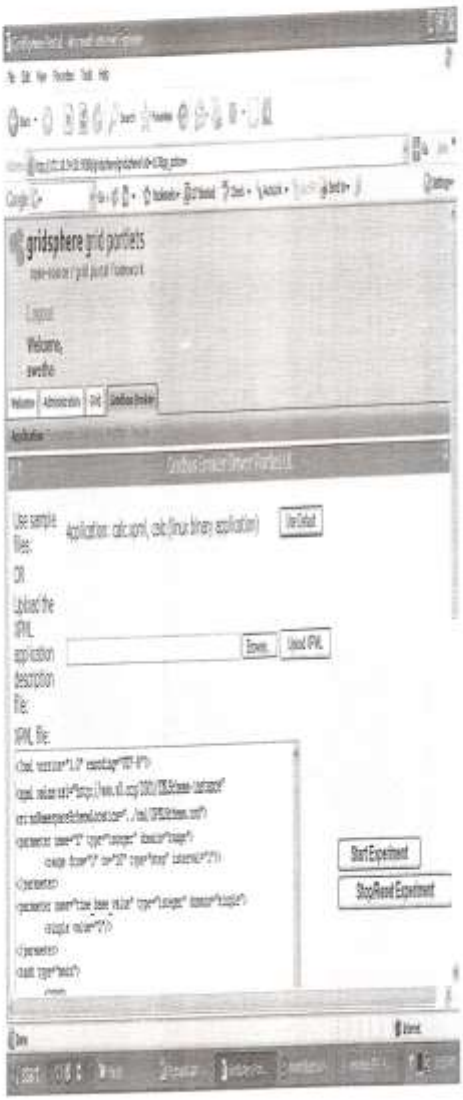


Figure 10. Application Portlet

5.3 Monitor Portlet

After submitting the job the user can check or monitor the status of the job through this portlet. It gives the status about each job and the resource on which the job is running. The status of the job could be whether the job is waiting or the job is waiting or the job is running or the job is done or the job has failed.

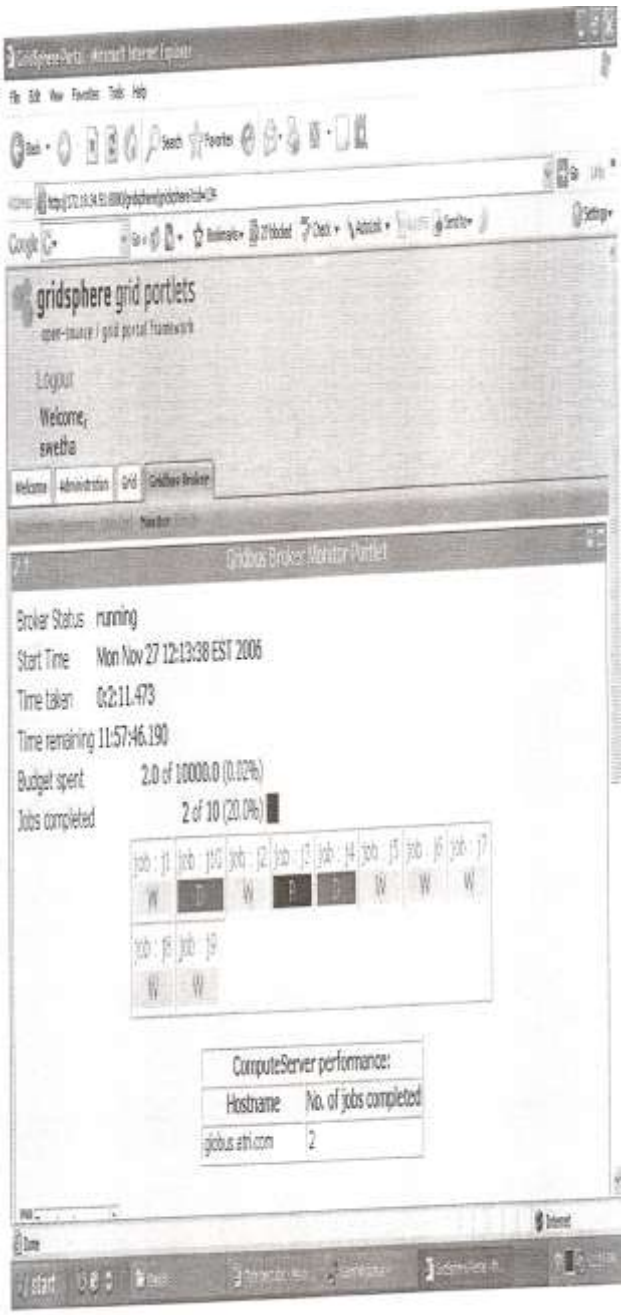


Figure 9: Monitor Portlet

Figure 11. Monitor Portlet

5.4 Result Portlet

Once the execution of all jobs are done the GridBus Broker returns the results as a link to the result file on the result portlet and then the user can download the output files and check for the results.

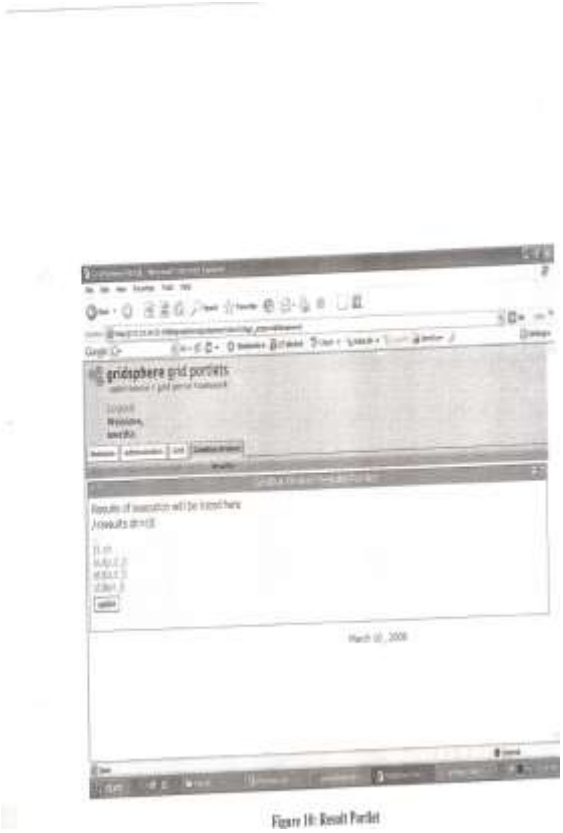


Figure 10: Result Portlet

Figure 12. Result Portlet

5.5 Login Page

A GridSphere Portal Login page is as shown below and this is the entry point for users and then users can start developing their own portals.

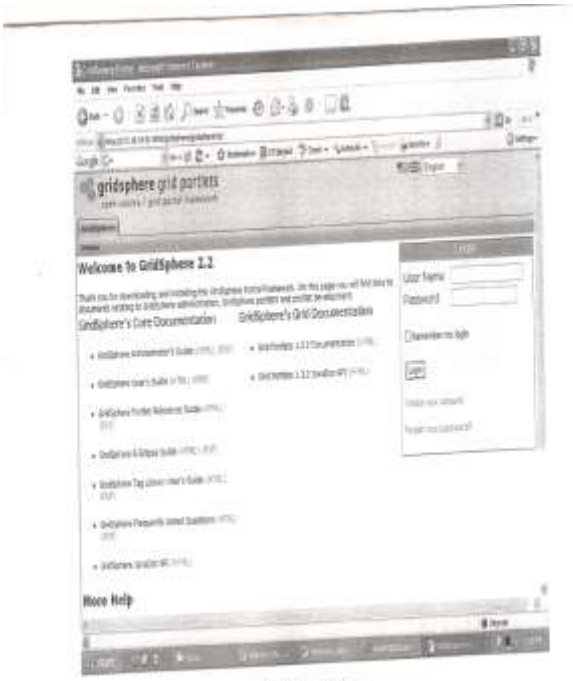


Figure 4: Login Portlet

Figure 13. Login Page

5.6 User QOS Portlet

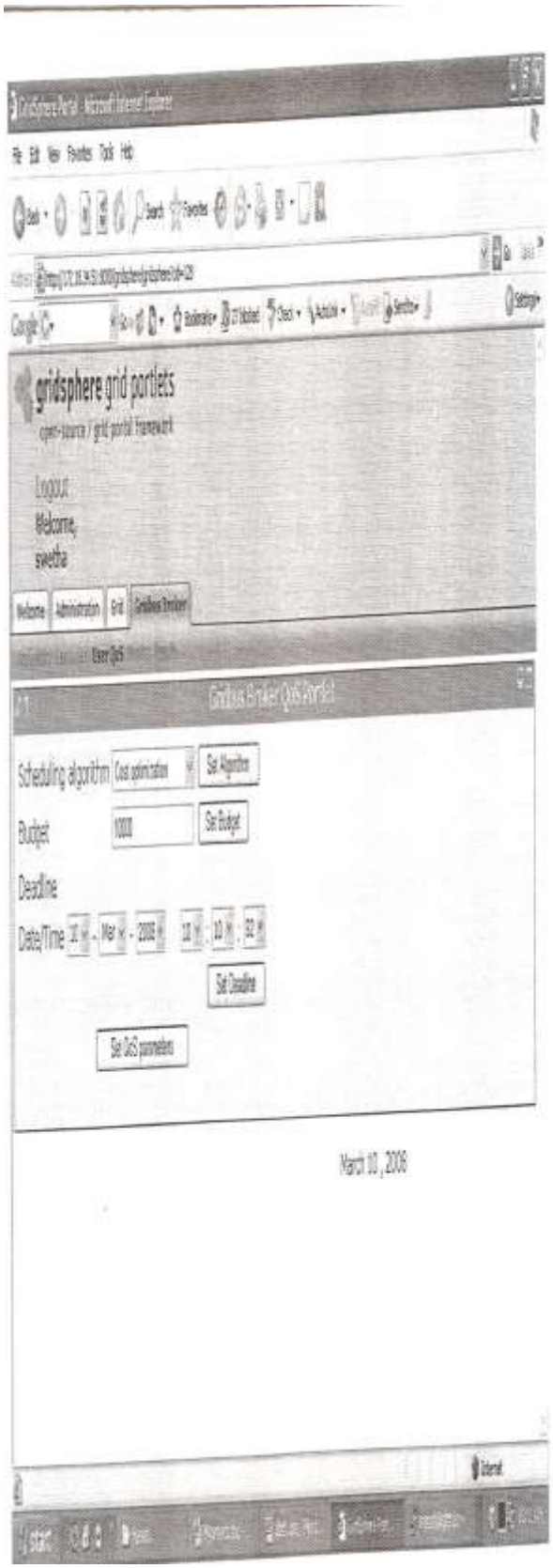


Figure 14. User QOS Portlet

5.7 Deadline

This is used to set the deadline for finishing the job. The deadline is set in the form of Date and Time.

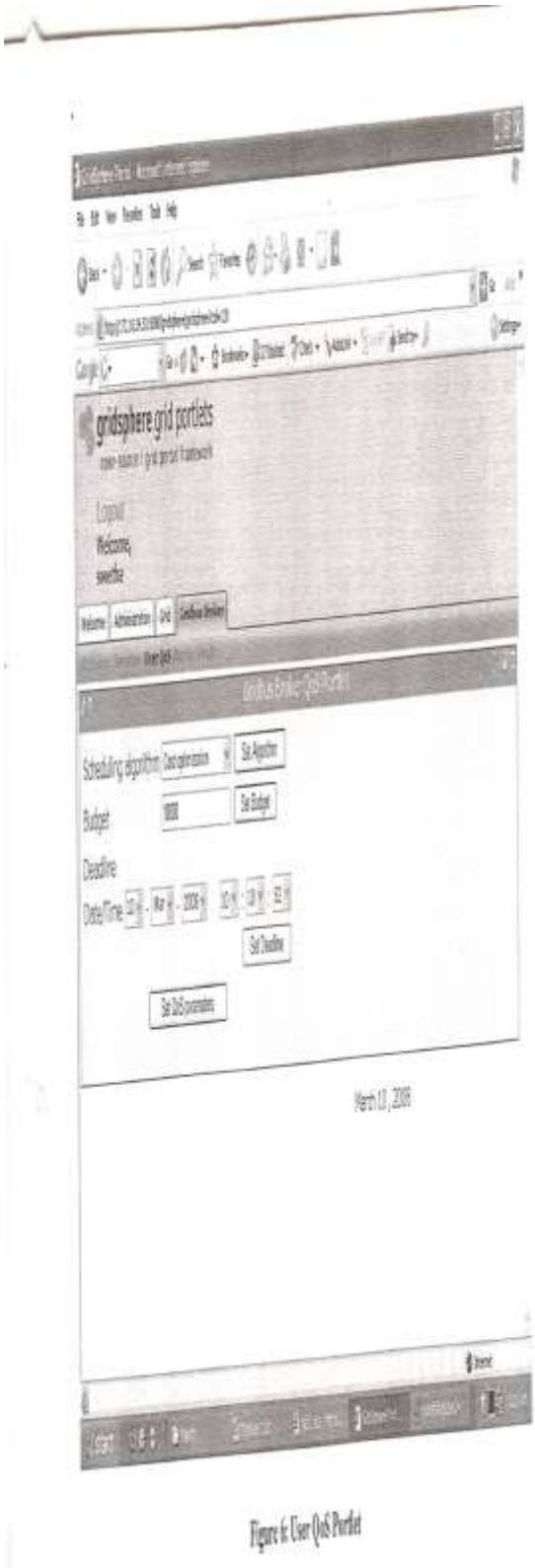


Figure 15. Deadline

6. CONCLUSION

This framework model attempts to explain the design of the GridSphere Portal, GridBus Broker and Globus.

7. ACKNOWLEDGMENT

This work was supported as Research Project implementation at "Osmania University", Hyderabad, India.

8. REFERENCES

[1] The Gridbus Grid Service Broker and Scheduler (2.0) User Guide http://www.cloudbus.org/reports/gridbus-broker-guide2.pdf

[2] http://phditsiamuresearch.pbworks.com/f/017.pdf

[3] http://www.rimtengg.com/coit2007/proceedings/pdfs/69.pdf

[4] http://gridcomputers.blogspot.com/

[5] High-Performance Computing On-Demand | The Scientist Magazine® http://www.the-scientist.com/?articles.view/articleNo/15271/title/High-Performance-Computing-On-Demand/

[6] InternationalReferredResearchJournal ISSN- 0975-3486 VOL,I http://www.ssmrae.com/admin/images/d378a351843796b8be2df2e4f0ef5f46.pdf

[7] An Integration of Global and Enterprise Grid Computing: Gridbus http://www.cloudbus.org/reports/gridbus-xgrid.pdf

[8] An Analysis of MIPS Group Based Job Scheduling Algorithm with ... http://ijcsi.org/papers/IJCSI-8-6-3-335-340.pdf

[9] Cluster and Grid Computing http://www.dcc.fc.up.pt/~ines/aulas/0910/MAPi/scheduling_gc.ppt

[10] J. Tourino, M. Martin, J. Tarrio, and M. Arenaz, « A grid portal for an undergraduate parallel programming cource,» IEEE Trans. Educ., vol.48, no. 3, pp.391-399, Aug.2005.

[11] A. Andronico, R. Barbera, A. Falzone, P. Kunszt, G. L. Re, A. Pulvirenti, and A. Rodolico, « Genius : A simple and easy way to access computational and data grids, » *Future Gener. Comput. Syst.*, vol. 19,no. 6,pp.805-813, 2003.

[12] R. Berlich, M. Kunze, and K. Schwarz, R. Buyya, P. Coddington, and A. Wendelborn, Eds., « Grid Computing in Europe : From research to deployment, » in *Proc. Australasian Workshop Grid Computing e-Research (AusGrid2005)*, Newcastle, Australia, 2005, vol. 44, pp. 21-27, ACS.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 517 - 521, 2013, ISSN: 2319–8656

FACTORS THAT AFFECT CLOUD COMPUTING ADOPTION BY SMALL AND MEDIUM ENTERPRISES IN KENYA

John Ngugi Makena
Jomo Kenyatta University of Agriculture and Technology
Institute of Computer Science and Information Technology
P.O. BOX 21403 -00100 Nairobi, Kenya.

Abstract: The number of Small and Medium Enterprises in Kenya have increased tremendously over the last ten years. They have played a profound role in providing employment to the population besides growing the Kenyan economy. Cloud computing is a new entrant to the technology arena which in form of Platform as a Service, Software as a Service and Infrastructure as a Service promises profound reduction in cost of operations in a business. It offers immense benefits as the business enterprise utilizes the pay per use model availed by the cloud service providers as per the needs of the business enterprise. This eliminates the need to purchase expensive software, development platforms and setting up complex ICT infrastructure. This is akin to renting what they need instead of purchasing and owning it.
However, the SMEs in Kenya have not taken up the cloud computing benefits to maximise their competitive advantage. This research paper focuses on the factors that affect the adoption of cloud computing by SMEs in Kenya. The research process involves a descriptive research design. The research findings have shown that technological, organisational and environmental factors have affected the adoption of cloud technologies.

Keywords: cloud computing, technological, organisational, environmental.

1. INTRODUCTION

Due to severe market competition and dramatically changing business environment, firms have been prompted to adopt various state-of-the -art Information Technologies to improve their business operations. (Pan and Jang, 2008; Sultan, 2010)

In modern technology arena, cloud computing has cut a great and specific niche in businesses. The use of information and communication technologies can improve business competitiveness and has provided genuine advantages for small and medium sized enterprises. (SMEs firms with one to 250 employees (DTI, 2004)), enabling them to compete with large firms (Swash, 1998; Bayo-Moriones and Lera-Lopez, 2007). In Kenya SMEs forms the largest block of employers. They provide the necessary and critical base for economic development. Competition for market share and great profit margins is cut throat. The growth of SMEs is compounded by their agility and adaptability to changing business models.

Cloud computing lowers IT costs and provides organisations with the people and expertise to create an integrated suite of software applications. (Subhankar, 2012)

Some of the promised benefits from cloud computing can be very appealing to SMEs, which need to maximise the return on their investment and still remaining competitive in an ever demanding business environment.(Yazn et al,2012).

The benefits of cloud computing adoption in developing countries have not been thoroughly exploited. If the SMEs adopt the robust services offered by cloud computing, they will have lowered the costs of operation because they would access business application softwares at a low cost. In view of this the research paper main objective is to contribute to a growing body of research on cloud computing by studying the determinants of cloud computing adoption by SMEs in Kenya.

The findings of this research will greatly influence the uptake of cloud services which could effectively and efficiently deliver services which could have otherwise been only accessible by large blue chips corporations and multinationals.

If SMEs have access to scalable technologies they could potentially deliver products and services that in the past only large enterprises could deliver hence flatten the competition arena. (Yazn *et al*, 2013)

While cloud computing has been discussed as a new technology development that can provide several advantages both strategic and operational to its adopters the cloud computing adoption rate is not growing as fast as expected (Goscinski and Brock,2010).

The SME sector in Kenya has been selected for this study because they are the largest providers of direct and indirect employment hence play a pivotal role in the economic growth of the country.

1.2. Towards cloud adoption

A survey on cloud computing awareness by Market Connections (2008) on US defence/ military and Federal government unearthed that cloud utilization is poised for rapid gains as awareness of cloud computing growth. The SME sector in Kenya has adopted various technology solutions that have enhanced their growth. ERPs and other integrated business software systems have been used by big corporates though the SMEs have not managed to access them.

Before any organisation can use cloud computing, it must be aware of existence of such a technology, what it is used for and where it can be applied.(Kiiru,2011).The CEOs of various SMEs in Kenya have adequate awareness of the benefits of technology in their organisations. However, only a few know the benefits of cloud computing beyond the basic description.

Many have adopted a wait and see attitude as far as adoption of cloud computing is concerned.

According to Ellison (2010), the concept of cloud computing has aroused interest in the enterprise but it is also clear that businesses are testing their options to decide whether they will adopt this technology.

Though there has been increased awareness by cloud services providers on organisational benefits of cloud computing, the rate of adoption is significantly low. A survey conducted by Daniel (2010) titled 'Mobile Technology and Business' in the United States revealed that 95% of the respondents were generally aware of the cloud computing concept and believe that its role and significance will increase in the next five years. There are various worries from the SMEs about cloud services adoption which contribute to the slow adoption in Kenya.

There have been inadequate research on utilization of technology for business advantage by SMEs in Kenya; various managers expressed concerns of security as a major drawback to the cloud services adoption.

According to Mime Cast (2010), majority of organisations (51%) in UK and USA using some form of cloud computing service and the levels of satisfaction amongst these organisations is high.

1.3. How ready are SMEs to adopt cloud computing?

"Computing services on-demand" is gradually modifying the way information system services are developed, scaled, maintained and paid for. (Yazn *et al*, 2013).The SMEs readiness to adopt new technology is a determinant of how they will adopt various forms of technology.

They more oftenly have a strong clinch to their former and current existing technologies such that adopting new ones may be difficult decision to make. This is partly due to the perceived losses of existing technology.

From a study carried out in India(Tripathi,2009)with the aim of finding out whether organizations are willing to adopt the technology, the results revealed that cloud computing has not gone past the awareness phase. The research showed that many decision makers are not aware of IaaS, PaaS and SaaS and their differences between public, private and hybrid clouds.

Many SMEs do not want to try a new technology which they are not knowledgeable of due to presumed cost of acquisition and perceived cost of failure. Most of the research and surveys on cloud computing have largely been done in the developed world other than Africa. Hence, the SMEs have scanty references whenever they may need relevant information concerning utilization of cloud services and the rate of success in Africa.

2. TOE FRAMEWORK

The TOE is an organisational level theory that provides a framework for technology adoption by an enterprise. It is a theoretical model for cloud computing diffusion needs to consider the weakness in the adoption and diffusion technological innovation which are caused by the specific technological, organisational and environment contexts of the firm. (Chinyao and Mingchang 2011).

Technological context represents the internal and external technologies related to the organisation. The technologies may involve equipment or practice. Organisational context is related to the resources and the characteristics of the enterprise. Environment context refers to the area in which a firm conducts its business which can be related to surrounding elements such as presence of related businesses, competition and availability of other technologies. All the three contexts present both constraints and opportunities for technological innovation and adoption. (Tornatzky and Fleischer, 1990) which influence the technology adoption by firms.

The TOE studies have shown that the following three features influence cloud computing adoption. Technological context(relative advantage, complexity and compatibility);Organisational context(top management support, firm size and technological readiness) and Environmental context(competitive and trading partner pressures).Therefore, TOE provides a clear theoretical basis and a consistent empirical support and the likely outcome of new technology adoption in a firm(Khan and Chau,2001;Zhu et al,2004;Shirish and Teo,2010).Since new technology adoption is a complex issue for many SMEs, this research utilizes the TOE framework as framework that determines the factors that affect cloud computing adoption by SMEs in Kenya.

3. METHODOLOGY

3.1 Research design.

The research objective is to study the factors that affect cloud computing adoption by SMEs in Kenya. This involves a qualitative study of the factors and their extent of influence based on the TOE framework.

3.2. Sampling and data collection

A total of 220 interviews with ICT officers and business owners distributed proportionately were carried out. However, a total of 260 contacts were made which resulted to 202 interviews. This resulted to a response rate of 78%.Among the 202 respondents, 55% were ICT officers who were mandated with making key decisions in the firm while 45% were owners of the business enterprises. Questionnaires were used to collect primary data through interviews. The respondents were owners and ICT officers. The data obtained was analysed using descriptive statistics.

4. RESULTS AND DISCUSSION

This research was greatly guided by the TOE framework. The TOE framework has been credited and proposed by Shirish and Teo (2010); Lin and Lin (2008) as precise approach towards analysing IT adoption by firms.

From the findings of the research, it emerged that up to 70% of the respondents were aware of the presence of cloud computing technology in the modern business environment. About 50% were aware of the benefits of the technology to their specific business model. However, 30% of the respondents never expected any benefits of adopting the technology.25% of the respondents were willing to switch

their current technology to cloud platforms while 60% were not willing to abandon their current technology and adopt cloud computing model.50% expressed their desire to adopt cloud computing technology gradually in the next five years.

4.1 Environmental context

Trading partner and competitive pressures played a significant role in determining whether to adopt cloud computing. From the research findings, it was observed that environmental context was influenced by the nature of the business. For example effects of environmental context on firms involved in manufacturing were different from those involved in services industry. This is because firms on same business activity were competing with each other. If one has not embraced technology, then others did not find solid reason to adopt cloud computing. However, those in sectors that technology was key were willing to adopt cloud computing so as to be more competitive. Their customers also put pressure on these firms causing a positive effect on the technology adoption. However, these pressures were not leading to a compulsory acquisition of technology since the customers received the services they wanted using current set up. These findings show that firms are aware of technological needs to enhance proper delivery of services but they choose not to shift to that technology. However, when they are pressurised by competition they aggressively adopt such technologies swiftly. This revelation is consistent with earlier studies from Chong and Ooi (2008) and Oliveira and Martins (2010) and implies that when firms are faced with intense competition, they tend to implement changes more aggressively. For example Wal-Mart, which requires its partners to either adopt radio frequency identification or lose their business (Chong and Ooi, 2008).

From the findings, it is clear that firms need to be made aware of specific benefits of adopting cloud services in the areas of business operations. This will result to a profound plan for adoption since the businesses will be aware of cost implications and expected profits.

4.2 Organisational context

The top management support was found to be a significant factor in the cloud adoption. This research found out that top managers were not willing to invest resources to new technology when there is an existing and working technology. However, some managers unwillingness to adopt was as a result of scanty knowledge they have about cloud computing. They cited that it is a new technology hence a‘wait and see” approach was the best for them. These findings are consistent with those of Lertwongsaien and Wongpinunwatana (2003) and Ramdani and Kawaleki (2007, 2008) who found that without top management support, SMEs are less likely to adopt new technologies.

The technology readiness of the firm also affected the adoption of cloud computing technologies. Firms with fewer employees with adequate ICT knowledge were generally not ready with adoption as opposed to those whose employees were having adequate ICT skills. For example technology based firms were willing to adopt the technology.

Small firms also expressed their desire to adopt the technology while larger ones were less willing to change. This could be attributed to the fact that decision making is easier in a small than the long and tedious protocol based decision making process in large organisations. This is consistent with Chinyao and Mingchang (2011) that organisational competency and readiness may help to leverage existing information technology applications and data resources across key processes along the value chain when the firms embeds the cloud computing service.

4.3 Technological context

Benefits that a firm expects to accrue from a certain technology motivates it to adopt it because employee appreciation of the relative advantages of the adopted system to raise work efficiency. Cloud computing has an advantage of increased efficiency in provision of services. The relative advantage of cloud computing services implementation could improve speed of business communications, efficiency of coordination among firms, customer communications and access to market information mobilisation.(Ambrust et al,2010).The findings of this research showed that the respondents were aware of the relative advantages of cloud computing. However, their inadequate knowledge about the cloud concept proved to be a major drawback towards making decisions pertaining adoption. Though the relative advantage was clear to most of the respondents, they cited cost of implementation is high. Therefore, this is consistent with Teo et al (2009) who cited that cost of implementing new technology systems, the cost of the systems themselves can be comparatively high and often represent a major barrier to their adoption.

The complexity of the proposed cloud services and compatibility were also points of concern by many respondents. However, these were not primary factors to consider when adopting such technology. The findings showed that several respondents believed that there will be seamless compatibility with their current technology or with few manageable modifications of their current systems. These findings were inconsistent with Oliveira and Martins, (2010).This findings could be due to the fact that many respondents were ICT officers who were generally conversant with operations of networks and databases. They may have been involved in previous compatibility issues and complexity constraints involving other technologies. However, this does not translate to adequate understanding of cloud technology like how to set up a private cloud. Therefore, tough complexity and compatibility did not appear as a general point of concern, it could still pose a challenge in adoption of cloud computing.

5. CONCLUSION

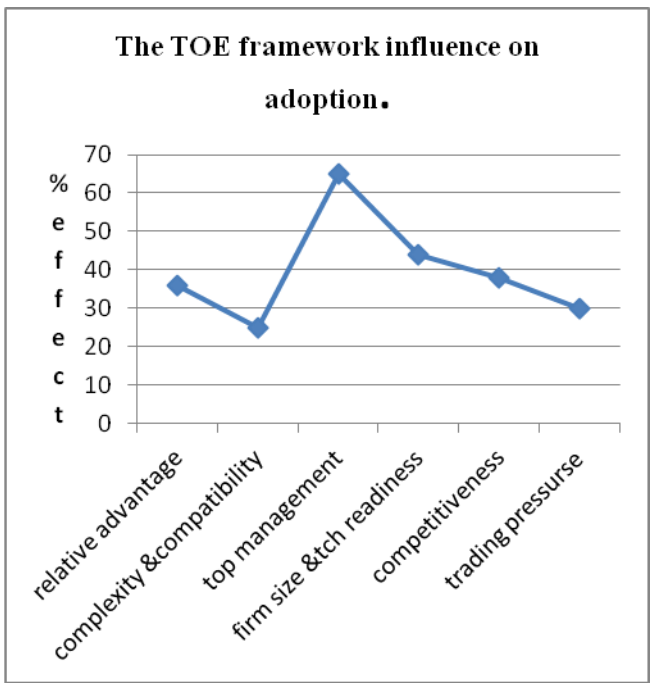
The dominant paradigm in studying IT adoption involves identifying contingency factors that facilitate or hinder the adoption decisions in an organisation.(Fichman,2004;Troshaniet al,2011).Though adoption of new technology is a complicated affair, the different factors in technological, organisational and environmental context vary across different innovations.(Bakers,2011).

The cloud service providers have an enormous task of creating awareness among all players in the business environment so that they can critically assess their internal structures and review them in readiness to adopt the new technology. Cloud computing is a relatively new phenomenon which have great potential of decreasing business operation costs and increase their efficiencies. The TOE framework has

provided an elaborate assessment criterion of factors influencing cloud adoption and it has shown that firms need to firm up their organisational, technological and environmental contexts so as to readily incorporate the new technology.

According to Yazn et al(2011) the nature of cloud computing should offer enough scope to generalize findings beyond the geographical area of study and to SMEs in other regions and even countries as cloud computing transcends boundaries and regional ICT infrastructure is not considered a major obstacle for the adoption process. The SME sector in Kenya has to position itself and align their objectives towards adoption of cloud computing as a new technology entrant in the market. The adoption of computing services on demand through a pay per use model will provide the SMEs sector with a wide pool of resources that befits their business model at a lower cost. Since the goal of any SMEs is to maximise profits, the resources saved after adopting cloud computing will be useful in expansion, research, improving the welfare of their employees and provision of quality services to customers.

In Kenya, technology advancements have fairly penetrated various markets. Therefore it will not be an uphill task to get personnel with technological prowess and appropriate technology base of running cloud applications over large networks. In future, further research could build on this study and find out qualitatively and quntitavely factors that affect cloud adoption by specific industries. Various stakeholders like cloud service providers, policy makers in Government should also be include in future research. SMEs sector in other countries could also be studied since all over the world the SME sector play profound role in economic growth. Therefore, their adoption of cloud computing and other relevant futuristic technologies will continuously position them in fast lane of growth and prosperity.



6. ACKNOWLEDGEMENTS

I humbly thank all professionals in the technology and SME sectors for their profound input in this research. I also thank all my colleagues and family for their tireless support in the process of this research.

7. REFERENCES

[1] Buyya R., C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Gener.Comput. Syst., vol. 25, pp. 599-616, 2009.

[2] Cena F., Farzan R., Lops P., Web 3.0: Merging Semantic Web with Social Web,

Proceedings of the20th ACM conference on Hypertext and hypermedia, HT'09, June 29–Jul1, 2009, page 385.

[3] Dorey P. G. and A. Leite, "Commentary: Cloud computing – A security problem or solution? "Information Security Technical Report, vol. 16, pp. 89-96, 2011.

[4] Erdogmus H., "Cloud Computing: Does Nirvana Hide behind the Nebula?" IEEE Software., vol.26, pp. 4-6, 2009.

[5] Kiiru, W. K. (1991). A review of the institutional lending to the jua kali and small enterprise sector in Kenya. Geneva: International Labour Organisation.

[6] King, K. & McGrath S. (2002) Globalisation, Enterprise and Knowledge:

Educational Training and Development, International Review of Education, Vol. 50(1).

[7] Khajeh-Hosseini A., D. Greenwood, and I. Sommerville, "Cloud Migration: A Case Study of

Migrating an Enterprise IT System to IaaS," presented at the Proceedings of the 2010 IEEE 3rdInternational Conference on Cloud Computing, 2010.

[8] Krutz, R.L. and Vines, R.D. (2010), Cloud Security: A Comprehensive Guide to Secure Cloud Computing, Wiley, New York, NY.

[9] Lillard T. V., C. P. Garrison, C. A. Schiller, and J. Steele, "Chapter 12 - The Future of Cloud Computing," in Digital Forensics for Network, Internet, and Cloud Computing, Boston: Syngress, 2010, pp. 319-339.

[10] Longenecker, J. G., Petty, C. W., Moore, J. W. and Palich, L. E. (2006). Small Business Management, An entrepreneurial emphasis. London: Thomson South Western.

[11] Marston S., Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing – The Business perspective," Decision Support Systems, vol. 51, pp. 176-189, 2011.

[12] Moyer, C.M. (2011), Building Applications in the Cloud: Concepts, Patterns, and Projects, Addison-Wesley Professional, Reading, MA.

[13] Saul Berman L. K.-T., Anthony Marshall and Rohini Srivathsa, "The power of cloud-Driving Business model innovation," IBM Institute for Business Value, Feb 2012.

[14] Shayan J., A. Azarnik, S. Chuprat, and M. Zamani, "Identifying security risks of exploiting Cloud computing in Educational environment," 2012.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 522 - 525, 2013, ISSN: 2319–8656

Classification of Breast Cancer Samples Through Using the Artificial Bee Colony Algorithm

Mahsa Nazarian
Department of Computer Science and Engineering
Khouzestan Science and Research Branch,
Islamic Azad University
Ahwaz, Iran

Mashala Abbasi Dezfouli
Department of Computer Science and Engineering
Khouzestan Science and Research Branch,
Islamic Azad University
Ahwaz, Iran

Ali Haronabadi
Department of Computer
Islamic Azad University,
Central Tehran Branch
Tehran, Iran

Abstract: The algorithm of artificial bee colony has been widely applied in optimization issues. The pattern of this algorithm has been derived from the intelligent behavior of bees in the nature. The aim of this paper is to present a model for classification of binary problems through using data mining techniques, and on the basis of artificial bee colony algorithm. In this paper, breast cancer data that has been presented by Wisconsin University to investigate and evaluate the proposed method has been used and applied. With regard to the obtained results, it has been shown that this algorithm has a considerable and important function.

Keywords: breast cancer, data mining, ABC, artificial bee colony, classification

1. INTRODUCTION

By using data mining process, knowledge can be obtained via input data set through various steps and procedures. Using data mining techniques widely in todays different sciences and problems have made it possible for researchers to explore the new concepts and patterns. In the past, exploration of these problems and issues was impossible. Double problems in classification refer to those problems and issues in which data is classifies into two groups and classes. Data mining includes two steps, namely pre-processing and pattern diagnosis. In pre-processing step, the considered characteristics and specifications are obtained via data, and the result is data without any errors. Pattern diagnosis step involves various algorithms for classification of data obtained from the pre-processing step. The algorithm of artificial bee colony is an algorithm based on the crowd. This algorithm has been increasingly considered because understanding and applying this algorithm is easy. This algorithm has been used in different fields such as optimization. In this paper, a model based on the artificial bee colony algorithm has been presented for exploring knowledge via data base related to breast cancer. In fact, the algorithm of artificial bee colony has been used and applied in the field of data classification. In this paper, in pre-processing step, the related data is firstly normalized. In the second step, through using the model based on the artificial bee colony algorithm, data is classified into two pre-defined classes.

2. ARTIFICIAL BEE COLONY ALGORITHM

Artificial Bee Colony Algorithm has been presented in 2005. Since understanding and applying this algorithm is easy, it has been widely used in various fields of optimization. The results of using this algorithm has been compared with other widely used methods such as genetic algorithm [4], DE (differential Evolution) [5], and PSO (Particle Swarm Optimization) [6]. This algorithm involves three parts: food source, worker bee and non-worker bee. There are two kinds of non-worker bees, namely pioneer bees and supervisor bees. The worker bees refer to those bees finding the source of food, so the number of theses bees equals to the number of food sources. In addition, the number of food sources equals to the number of problem solutions. The pioneer and supervisor bees try to find new food sources. The basis of finding food sources is random exploration. The pioneer bees find the food sources randomly, while the supervisor bees evaluate and investigate the quality. This algorithm can be defined on the basis of the following equations:

$$V_{ij} = X_{ij} + \phi_{ij}(X_{ij} - X_{kj}) \quad (1)$$

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (2)$$

In equation (1), P_i is the possibility of selecting ith food source. fit_i refers to the value of ith food source, and SN is the number of food sources. In fact, SN equals to the number of possible solutions [10].

In ABC algorithm, artificial bees find new solutions locally. This exploration has been shown in equation (2). V_{ij} is the location of the new food source. Both X_{kj} and X_{ij} are old food sources. I this equation, $j \in [1,2,...,D]$ and $k \in [1,2,...,SN]$ are considered. K and j values are not equal, and are selected randomly. D equals to the number of optimization parameters. ϕ is a random number [-1 and 1]. The important point is that, in selecting food source, ABC algorithm selects the sources whose quality equals to the prior source or is more than it. This means greedy selection. After choosing the new source of food, that source is added to the

memory. There are two other parameters in this algorithm, namely the limitation of food source location and the repeated number of food source exploration cycles [10].

3. LITERATURE

Osmar and his colleagues presented several methods for diagnosis of tumor in digital mammography in which two determining techniques, namely neural network and association rules, have been used for anomaly diagnosis and classification. In both methods, the precision rate of classification is more than 70 percent. 332 mammography images were used, and they were classified into three groups, namely normal tumor, benign tumor and malignant tumor. In addition, abnormal issues are classified to six groups, and it depends on the degree and kind of its growth. Their research involved several steps such as image reception, image correction, exploring symptoms and signs, and classification. They used apriori algorithm in order to derive association rules from the signs and symptoms [11].

Dr. Rani classified the related data of medicine through using neural networks. He used the model of feed forward neural network and the algorithm of back propagation learning. He also considered the strategy of parallel neural network. In his research, 699 samples and 10 specs were used. According to the results of classification, it has been reported that success percentage of the proposed method was 96,6 [13].

Anuncias and his colleagues presented a data mining approach for detection of high-risk breast cancer groups [3].

Abdelghani Bellaachia and his colleagues tried to predict breast cancer survivability using three data mining techniques, namely Naïve Bayes, neural networks and decision-making tree. It has been reported that the success percentage of Naïve Bayes neural networks and decision-making tree were respectively 84,5 and 86,5 and 86,7 [7].

Chang and his colleagues tried to compare three data mining techniques with genetic algorithm, and they focused on the function of artificial intelligence and data mining. They presented a model for predicting breast cancer [12].

Amin Einipour and his colleagues presented a fuzzy method on the basis of ants colony for breast cancer diagnosis. The algorithm of ants colony has been derived from the behavior of ants searching the food in the shortest direction. In this research, the authors presented a method in order to classify the samples into two groups, namely cancroids and un-cancroids samples, and they used fuzzy rules and method as well as ants colony algorithm.[1]

Ping zhang and his colleagues explored Bayesian networks for breast cancer detection. These tools are used by radiologists [8].

Gupta and his colleagues investigated breast cancer diagnosis and prognosis. Also, they explored the medicine data and their classifications.[9]

4. USED DATA

In this paper, Wisconsin data set related to breast cancer samples has been used and applied in order to evaluate and investigate the proposed method. This set involves 699 records and 9 specs. In this data set, there are records having incomplete data. At first, incomplete records must be deleted. The number of these records is 16. When these records are deleted, 683 records remain.

5. THE PROPOSED METHOD

5.1 Pre-processing

As it was mentioned above, incomplete and non-numerical data must be deleted, and should not be considered in data set; therefore, in the first step, data records must be selected so that these data will enter the next step. Here, we consider 450 records as the training data. These records are used for exploring the rules as well as evaluating them. We consider remaining 233 records as the experimental records. After exploring and obtaining rules, all 683 records will be classified.

5.2 The Pattern of rules

Each classification rule involves two parts in classification, namely condition and result. The rule structure of this method is as follows:

```
Struct RuleSet{
        Double *lowb;
        Double *upb;
        String ClassName;
        Double *fitval;
}
```

lowb is the sign of minimum values of each spec. upb is the sign of maximum values of each spec. ClassName refers to the name of the class to which the rule is related, and fitval stands for the value of the rule fitness.

5.3 Evaluation Function

In order to evaluate the fitness of each rule, the following evaluation function has been proposed.

$$FV = \frac{TP}{TP+FN} \times \frac{TN}{TN+FP} \text{ [10]} \qquad (3)$$

In this equation, FV is the rule fitness value. TP refers to the number of records that have also been applied in rule condition, and their class is as same as the class rule. TN stands for number of records that have not been applied in rule condition, and their class is not as same as the class rule. FN is the number of records that are not used in rule condition, but their class is as same as the class rule.

5.4 Rule Exploration and Extraction

In this step, we explore the rules through using random numbers. Here, we create two random minimum and maximum values by using equations (4) and (5). The created random numbers neither should be too great nor too small. For instance, if the maximum value is too great, all records will be applied.

$$LB = f - \phi_1(F_u - F_l) \text{ [2]} \qquad (4)$$

$$UB = f + \phi_2(F_u - F_l) \text{ [2]} \qquad (5)$$

In these equations, LB is random minimum value, and UB is random maximum value. ϕ_1 and ϕ_2 are two random numbers, and their values range from 0 to 1. f stands for the current record value. F_u is the maximum value of that spec, and F_l refers to maximum value of the spec. since the ranges of specs in the investigated data set are between 0 and 10, we changed the above mentioned equations to the following equations:

$$LB = f - 9 \times \phi_1 \quad (6)$$

$$UB = f + 9 \times \phi_2 \quad (7)$$

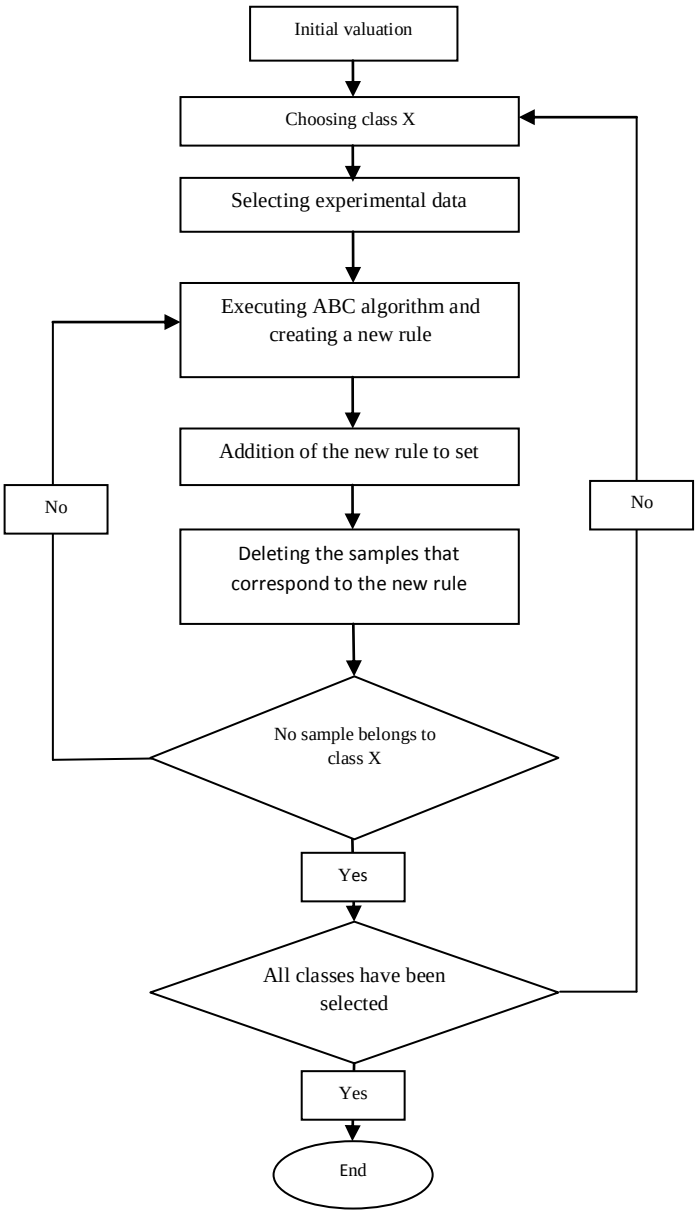


Figure.1 flow chart of rule extracting

The flow chart of rule extraction has been shown in figure (1). The most important part in classification is rule extraction and exploration. since proper rules cannot be found, the efficiency of the method will be reduced. In this algorithm, rules are considered as the solutions or food sources, and they should be evaluated after extraction and exploration. In rule evaluation, the value of rule fitness is computed via equation (3), and the rule structure is stored in the memory. At the end of this step, there are many rules in the memory.

5.5 Rule Selection

As it was mentioned before, there are many rules in the memory. In order to use and apply these rules in classification, the best rule must be selected and considered. The criterion of selecting the best rule is fitness value. Through selecting the rule with maximum fitness value for each class, we consider data classification**.**

5.6 Selecting dominant Class

After selecting the best rule for each class, we select the class whose rule involves maximum fitness value. With regard to data available in WBCD set, the class whose distinguishing rule has the maximum fitness value is considered as benign class.

5.7 Classification

After selecting the rule related to dominant class, we try to classify data. In this step, each sample applied in rule condition in located in benign class. If a sample is not used in this rule, it will be located in malignant class, since there are two kinds of classes. This procedure will continue until all data are completed.

6. COMPARING THE METHOD WITH OTHER METHODS

In this paper, we compare the rules obtained from executing this algorithm in data related to breast cancer with other methods. In order to evaluate this method, the above mentioned algorithm is executed 10 times, and in each step, precision rate will be specified in diagnosis of samples. The results of this step have been demonstrated in table (1). The average precision rate of classification is 96,5% in the proposed method. The average precision rate of the proposed method and other methods has been shown in table (2). In order to use and apply the proposed method, visual studio, version 2005 and C# have been used. In addition, in order to compare this method with other methods, Weka software, version 3.6, has been used.

Table 1. The Results of the Proposed Algorithm Repetition in Breast Cancer Diagnosis

Repetition Stage	Percent of diagnosis precision	The number of correct samples	The number of incorrect samples
1	95.91	655	28
2	96.12	656	27
3	96.58	660	23
4	96.59	660	23
5	96.69	660	23
6	96.33	658	25
7	97.25	664	19
8	95.92	655	28
9	96.84	661	22
10	96.79	661	22
Average	96.5		

Table 2. Comparing the Proposed Algorithm with Other Methods

algorithm	Percent of diagnosis precision	The number of correct samples	The number of incorrect samples
The proposed method	96.5	659	24
Naive Bayes	96.48	659	24
Neural Network	95.17	650	33
decision tree-C4.5	95.75	654	29
Support Vector Machine	97.07	663	20

7. CONCLUSION

In this paper, the algorithm of artificial bee colony has been used for classification because this algorithm is a new, simple and efficient algorithm in terms of optimization. Classification is very important for data processing in data mining. In addition, breast cancer is one of the most prevalent cancers, and its growth can be prevented through on- time diagnosis and prognosis. In this paper, we have proposed a method on the basis of artificial bee colony algorithm; therefore, diagnosis of this illness will be possible through using data related to breast cancer samples. The results of this method have been compared with other methods of breast cancer diagnosis. It should be mentioned that diagnosis precision of this method has been considered as medium in comparison with other methods. This method is simpler than other methods, and its concepts are more understandable. In the future, we will try to increase the success percentage of the proposed method.

8. REFERENCES

[1] Amin E. 2011 A Fuzzy-ACO Method for Detect Breast Cancer Global Journal of Health Science : Vol. 3, No. 2 (October 2011)

[2] Afizi Mohd, Mujahid Ahmad, Zaidi Ahmad, "Artificial Bee Colony based Data Mining Algorithms for Classification Task", Modern Applied Science,Vol 5, No 4, 2011

[3] Anunciac Orlando ¸ Bruno C. Gomes, Susana Vinga, Gaspar Jorge, "A Data Mining Approach for the detection of High-Risk Breast Cancer Groups", 2004

[4] B. Basturk, D. Karaboga, An Artificial Bee Colony (ABC) algorithm for numeric function optimization, in: IEEE Swarm Intelligence Symposium, Indiana, USA,2006.

[5] B. Basturk ,D. Karaboga, Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems, LNCS: Advances in Soft Computing:Foundations of Fuzzy Logic and Soft Computing, vol. 4529, Springer Verlag, p. 789–798, 2007

[6] B. Basturk, D. Karaboga, A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm, J. Global Optim. 39 ,171–459, 2007

[7] Bellaachia Abdelghani,Erhan guven, "Predicting Breast cancer survivability using Data MiningTechniques.", 2005

[8] Gadewadikar Jyotirmay, Ognjen Kuljaca, Kwabena Agyepong, Erol Sarigul, Yufeng Zheng, Ping Zhang, "Exploring Bayesian networks for medical decision support in breast cancer detection", 2010

[9] Gupta Shelly, Kumar Dharminder Dean, Anand Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", 2011

[10] D. Karaboga, "An idea based on honey bee swarm for numerical optimization",Technical Report-TR06, Erciyes University, Engineering Faculty, ComputerEngineering Department, 2005.

[11] Luiza Antonie Maria, Osmar R. Za¨ıane, Alexandru Coma, "Application of Data Mining Techniques for Medical Image Classification", 2001

[12] Wei Pin, Chang, Der-Ming, Liou , " Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data ",2007

[13] Rani K. Usha, "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique", 2002

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 526 - 529, 2013, ISSN: 2319–8656

Algorithm for P versus NP Problem on Sets by JEEVAN – KUSHALAIAH Method

Neelam Jeevan Kumar
Electrical and Electronics Engineering, JNTU Hyderabad
Rangashaipet, Warangal, Andhra Pradesh, India.

Abstract: P versus NP [1-2] Problems are one of the most important open questions in mathematics and theoretical computer science. Jeevan – Kushalaiah Method is a method to find the possible number of combinations between n-elements. This article explains about Algorithm to solve subset sum problem quickly and easily. The problems on subset sum problems perform arithmetic operations that can be calculated in terms of exponential time - polynomial time. Major part of the article deals with Class-P type problems which in be solved on a deterministic Turing Machine. This article is mainly prepared on the basis of an article in Wikipedia.

Keywords: Jeevan-Kushalaiah Method; Jeevan – Kushalaiah Algorithm; Time Complexity; Exponential Time; Polynomial Time;

1. INTRODUCTION

In computational complexity theory an answer to the P = NP [3-4] question would determine whether problems that can be verified in polynomial time, like the subset-sum problem, can also be solved in exponential- polynomial time. If it turned out that P ≠ NP, it would mean that there are problems in NP (such as NP-complete problems) that are harder to compute than to verify: they could not be solved in polynomial time, but the answer could be verified in polynomial time. The algorithm named as Jeevan – Kushalaiah Algorithm because it uses Jeevan-Kushalaiah Method. The Time taken to produce the output is in Exponential -polynomial time

$$\mathbf{T(n1) = EXPTIME = } \ 2^{\mathbf{poly(n1)}} \ \mathbf{.....(I.a)}$$

$$\mathbf{T(n1) = P = (2^{O(\log n)} = poly(n)) \quad ..... \quad (I.b)}$$

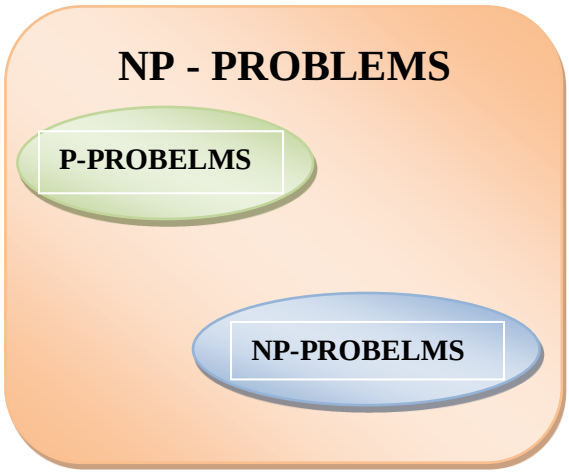


Figure.1: Diagram of complexity classes provided that P ≠ NP. The existence of problems within NP but outside both P and NP-complete, under that assumption, was established by Ladner's theorem

2. JEEVAN – KUSHALAIAH METHOD

Jeevan – Kushalaiah Method [5] is a method to know possible different combinations between n-elements (either elements are constants or variables).

Let there are n-elements $\mathbf{a_1, a_2, a_3 \ldots a_n}$

The possible number of Combinations are in sets with addition operation are, N

$N=\{0,[(a_1),(a_2),(a_3),\ldots.(a_2)],[(a_1+a_2),(a_1+a_3),(a_1+a_4),\ldots..(a_1+a_n),(a_2+a_3),(a_2+a_4),(a_2+a_5),\ldots (a_2+a_n),(a_3+a_4), \ldots. (a_3+a_n),\ldots.(a_{n-1}+a_n)],[(a_1+a_2+a_3),(a_1+a_2+a_4),(a_1+a_2+a_5), \ldots..(a_{n-2}+a_{n-1}+a_n)]\ldots\ldots. [\ldots.]\ldots\ldots. [(a_1+a_2+a_3+\ldots+a_n)]\}$

$N= \{\Theta_0 , \Theta_1, \Theta_2, \Theta_{n-1}, \Theta_n \}$

Where $\Theta_1 = [(a_1),(a_2),(a_3),\ldots.(a_2)]$ ..... $\Theta_n = [(a_1+a_2+a_3+\ldots+a_n)]$

$\Theta_n = [\phi_{n,1} , \phi_{n,2} , \ldots. \Phi_{n,p}]$

Maximum value of $\phi_{n,p}$ is

$$\mathbf{\Phi_{n,p} = {}^nC_p = \binom{n}{p} = (n!)/(n-p)!(p)! \ldots\ldots (II)}$$

$$\mathbf{\Theta_n = \sum_{i=1}^{\binom{n}{p}} \phi_{n,} \binom{n}{p} \ldots\ldots \quad (III)}$$

The possible number of Combinations, N

$$\mathbf{N = \sum_{i=1}^{n} \Theta_n \ldots.. (IV)}$$

3. JEEVAN – KUSHALAIAH ALGORITHM

To find Subset of Set is Predefined value, x by adding elements of Set. To find which subset combination has given predefined value by adding has to Jeevan-Kushalaiah Algorithm.

Jeevan – Kushalaiah Algorithm must satisfy three conditions on Sets before going to algorithm process.

Condition-1: Create Subsets with positive and negative numbers and name as P_n and N_n respectively

Condition-2: The sum all values of either set should not be less than minimum value of other set

Mod{$\sum P_n$} not < Mod{min[N_n]} or

Mod { $\sum N_n$ } not < Mod{min[P_n]}

Condition-3: Eliminate or remove or delete Θ_0 value from P_n and N_n

Condition-4: The set should not contain all same sign numbers

The algorithm steps:

Step-1: Start the Program

Step-2: Perform condition-1 on given set.

Step-3: Create possible number of subsets in P_n , N_n with equations (II),(III) and (IV)

Step-4:Check condition-2 if it fails display output as "NOT POSSIBLE' then go to step-10 otherwise go to step-5

Step-5: Set Z = 0, Where Z is the number of Predefined valued subsets.

Step-6: Add each subset of P_n (or N_n) with N_n or (P_n) and equate the result with predefined value.

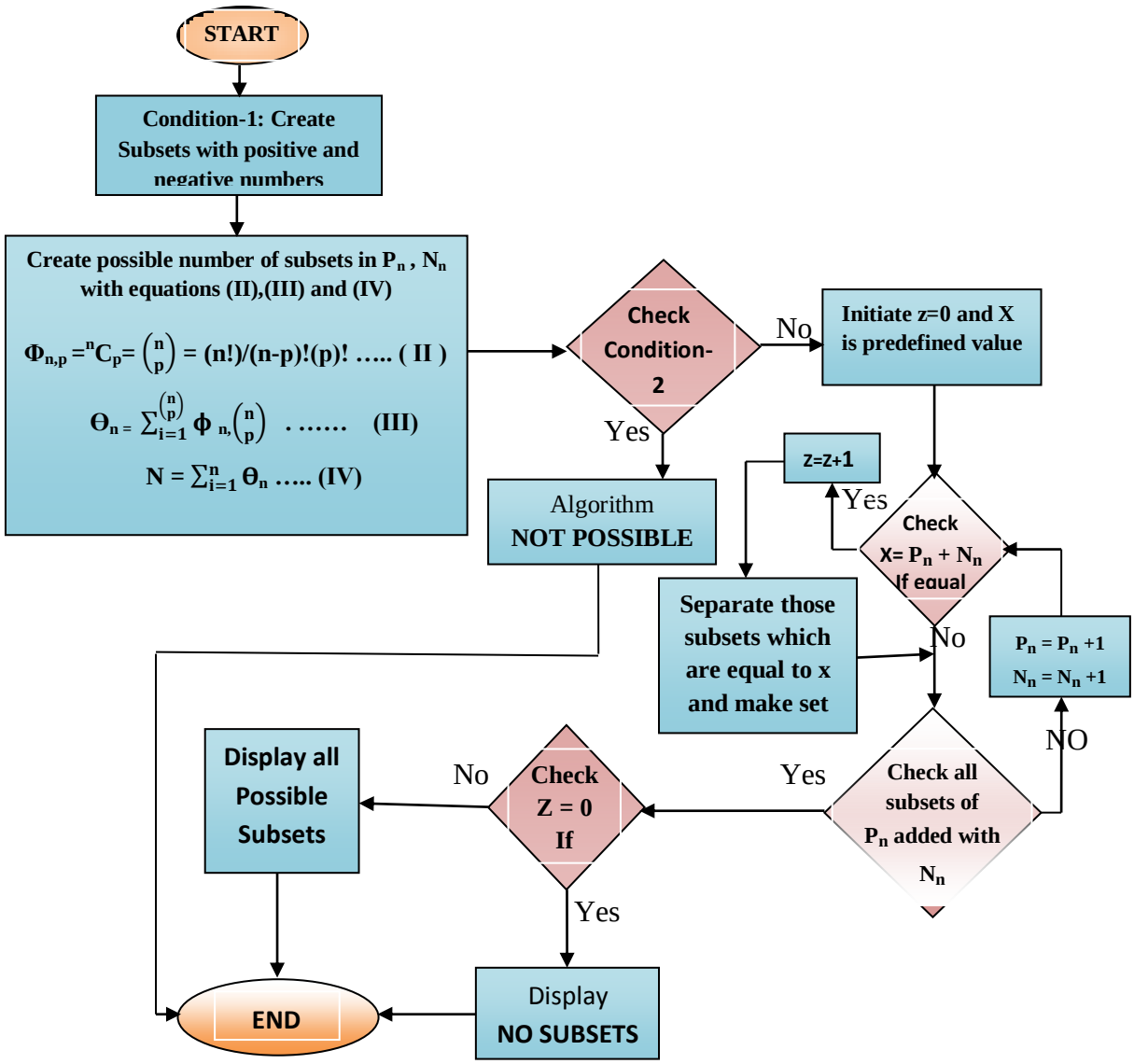
Step-7: Increase Z= Z+1when result is equal to X and Separate those subsets which are equal to x and make a set.

Step-8: Check all possible of combinations are done

Step-9: then check Z value if it is Zero display 'NO SUBSETS' otherwise "Display all Possible Subsets'. Then Go to Step-10

Step-10: Terminate (or END) the Program

4. FLOW CHART FOR JEEVAN – KUSHALAIAH ALGORITHM



- Number of iterations of algorithm = [no.of subsets of $\mathbf{P_n}$]*[no.of subsets of $\mathbf{N_n}$] **… (V)**
- The area of Jeevan – Kushalaiah Algorithm table is NxN (Where N is from equation (IV))

5. COMPUTATIONAL TIME

The total computational time[6-7] **T(n)** = EXPTIME(P) is directly proportionate to N_1*N_2(time taken for each addition)

Where EXPTIME = exponential time[8] for calculating subset elements of P_n and N_n
and P = polynomial time[9] for calculating arithmetic operation (i.e., addition) between two elements between P_n and N_n

N_1 = Number of subsets of P_n and N_2 is number of subsets of N_n

$$\mathbf{EXPTIME = 2^{poly(n1)} \ldots\ldots\ldots\ldots \text{(from equation I.a)}}$$

$$\mathbf{P= 2^{O(\log n2)} \ldots\ldots\ldots\ldots\ldots \text{(from equation I.b)}}$$

$$\mathbf{T(n) = EXPTIME(P) = (2^{poly(n1)})(2^{O(\log n2)}) \ldots\ldots\ldots (VI.a)}$$

The arithmetic time is directly proportional to $N_1 * N_2 = n_2$

$$\mathbf{T(n) \alpha EXPTIME(n_2)\ldots\ldots (VI.b)}$$

6. EXAMPLE

Does a subset of the set {−2, −3, 15, 14, 7, −10} add up to x=0 ?

Checking Jeevan-Kushalaiah Algorithm Conditions

Condition-1: Create Subsets with positive and negative numbers and name as P_n and N_n respectively

P_n = {15,14,7}and N_n = {-2,-3,-10}

Condition-2: The sum all values of either set should not be less than minimum value of other set

$$\mathbf{Mod\{\sum P_n\} \text{ not } < Mod\{min[N_n]\} \quad or \quad Mod\{\sum N_n\} \text{ not } < Mod\{min[P_n]\}}$$

Mod(14+14+7) > Mod(-2) and Mod(-2-3-10) <Mod(7), Condition – 2 is satisfied

Condition-3: Eliminate or remove or delete Θ_0 value from P_n and N_n

P_n = { Θ_1, Θ_2,Θ_3} and N_n = { Θ_1, Θ_2,Θ_3} : number of subsets, N = 7 from equations (I),(II) and (III)

P_n = { [15],[14],[7],[(15+14=29),(15+7=22),(14+7=21)],[15+14+7 = 36]} = {[15],[14],[7],[(29),(22),(21)][36]}

N_n = { [-2],[-3],[-10],[(-2-3=-5),(-2-10 = -12),(-3-10 = -13)],[-2-3-10 = -15]} ={[-2],[-3],[-10],[(-5),(-12),(13)],[-15]}

$\mathbf{P_n}$:Θ_1 = [(15),(14),(7)] , Θ_2 = [(29),(22),(21)] , Θ_3 = [36]; $\mathbf{N_n}$:Θ_1 = [(-2),(-3),(-10)] , Θ_2 = [(-5),(-12),(-13)] , Θ_3 = [-15]

By the algorithm flow chart equation – (VI):

Number of iteration = [no.of subsets of $\mathbf{P_n}$]*[no.of subsets of $\mathbf{N_n}$] = [7]*[7] =49

7. CONCLUSION

The manuscript explained and proposed about Subsets of sets which are equal to predefined value, X by adding. The elements of the Set are either Constants like Trigonometry, Logarithmic, Exponential Functions or Integers like K1, K2…. Any type of arithmetic operations like addition, subtraction, multiplication and division is performed on the sets. Many computer scientists believe that P≠NP and Author too believe. The main reasons for failing of Jeevan – Kushalaiah algorithm is conditions on sets and Computational Program for Set. Even though all of them are satisfied there may be a chance of getting solution with Z=0(zero) its mean the solution is cannot be determined by Computer program but can be calculated the paper work and major drawback is there is no Algorithm program in the world to perform operation on sets. The Example shows only for 6-element set as the elements of subset increases the complexity of the problem increases Exp (poly)..ly of computational time, T(n).

Table.1 : Jeevan – Kushalaiah Algorithm table to get Predefined value, X = Pn + Nn

X= P_n + N_n = 0	**X= P_n + N_n = 0**	**X= P_n + N_n = 0**	**X= P_n + N_n = 0**	**X= P_n + N_n = 0**	**X= P_n + N_n = 0**	**X= P_n + N_n = 0**
15 – 2 =13	14 – 2 =12	7 – 2 =5	29 – 2 =27	22 – 2 =20	21 – 2 =19	36 – 2 =34
15 - 3 =12	14 - 3 =11	7 - 3 =4	29 - 3 =26	22 - 3 =19	21 - 3 =18	36 - 3 =33
15 – 10 =5	14 – 10 =4	7 – 10 = -3	29 – 10 =4	22 – 10 =12	21 – 10 =11	36 – 10 =26
15 -5 =10	14 -5 =19	7 -5 = 2	29 -5 =24	22 -5 =17	21 -5 =16	36 -5 =29
15 -12 =3	14 -12 =1	7 -12 = -5	29 -12 =17	22 -12 =10	21 -12 =9	36 -12 =24
15 – 13 =2	14 – 13 =1	7 – 13 = -6	29 – 13 =16	22 – 13 =9	21 – 13 =8	36 – 13 =23
15 – 15 =0*	14 – 15 =-1	7 – 15 = -8	29 – 15 =14	22 – 15 =7	21 – 15 =6	36 – 15 =21

Subsets which are equal to **X= P_n + N_n = predefined value (i.e., 0)**

ANSWER is {15,-2,-3,-10}

8. REFERENCES

[1]. Cook, Stephen, "The **P** versus **NP** Problem". Clay Mathematics Institute. April 2000, Retrieved 18 October 2006

[2]. Cook, Stephen "The complexity of theorem proving procedures". Proceedings of the Third Annual ACM Symposium on Theory of Computing. 1971, pp. 151–158.

[3]. Ben-David, Shai; Halevi, Shai (1992). On the independence of P versus NP. Technical Report 714. Technion

[4]. Elvira Mayordomo. "P versus NP" Monografías de la Real Academia de Ciencias de Zaragoza 26: 57–68 (2004).

[5]. Neelam Jeevan Kumar, Neelam Kushalaiah, "JEEVAN-KUSHALAIAH METHOD TO FIND THE COEFFICIENTS OF CHARACTERISTIC EQUATION OF A MATRIX AND INTRODUCTION OF SUMMETOR", IJSER, pp 1553-1562, 4(8), ISSN 2229-5518, Aug-2013

[6]. Lance Fortnow, Steve Homer, "The Computational Complexity Column", NEC Laboratories America, Princeton, NJ 08540, USA.

[7]. Christos Papadimitriou , "Computational Complexity". Addison-Wesley. ISBN 0-201-53082-1. 1991, page 491.

[8]. J. M. Robson, "N by N checkers is Exptime complete". SIAM Journal on Computing, 1984, 13 (2): 252–267

[9]. Terr, David. "Polynomial Time." From MathWorld--A Wolfram Web Resource, created by Eric W. Weisstein.

[10]. Impagliazzo, R.; Paturi, R.; Zane, F. "Which problems have strongly exponential complexity?", Journal of Computer and System Sciences, 2001, 63 (4): 512–530

Wikipedia Links:

[1]. http://en.wikipedia.org/wiki/Millennium_Prize_Problems

[2]. http://en.wikipedia.org/wiki/P_versus_NP_problem

[3]. http://en.wikipedia.org/wiki/EXPTIME

[4]. http://en.wikipedia.org/wiki/Polynomial_time#Polynomial_time

Books:

[1]. Christos Papadimitriou , "Computational Complexity". Addison-Wesley. ISBN 0-201-53082-1. 1991, page 491

[2]. Schrijver, Alexander, "Preliminaries on algorithms and Complexity". Combinatorial Optimization: Polyhedra and Efficiency 1, 2003, Springer. ISBN 3-540-44389-4.

[3]. Carlson, James; Jaffe, Arthur; Wiles, Andrew, eds. (2006). The Millennium Prize Problems. Providence, RI: American Mathematical Society and Clay Mathematics Institute. ISBN978-0-8218-3679-8

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 530 - 538, 2013, ISSN: 2319–8656

Bayesian Network for Uncertainty Representation in Semantic Web: A Survey

Kumar Ravi
Department of Computer Applications,
S. Sinha College, Magadh University,
Aurangabad, Bihar, India

Sheopujan Singh
Department of Mathematics and Computer Applications,
S. Sinha College, Magadh University,
Aurangabad, Bihar, India

Abstract: Bayesian network is a probabilistic model to represent uncertainty available in knowledge base and using it tremendous works have been done to prove its relevance in uncertainty representation and reasoning using Bayesian inference. Probability can be used to represent uncertainty like prediction information, situational awareness, data and knowledge fusion etc in knowledge base to implement various real life situations. Various approaches based on description logic, object oriented, entity relational, and first order logic have been tried to represent uncertainty successfully. One of them is Multi-Entity Bayesian Network (MEBN) logic to represent probabilistic information and performing knowledge fusion in ontology, which is realized using PR-OWL (Probabilistic Web Ontology Language). This paper aims at giving an overall view, the work carried out so far to represent uncertainty with the help of Bayesian Network in semantic web and a list of works done using MEBN/PR-OWL for knowledge fusion or the representation of uncertainty in semantic web.

Keywords: Bayesian Network, Uncertainty, PR-OWL, MEBN, Semantic Web

1. INTRODUCTION

Bayesian network is widely used method for the representation of uncertain data and knowledge [1]. Bayesian network is also known as recursive graphical models, Bayesian belief networks, belief networks, causal probabilistic networks, causal networks, influence diagram and many more. Bayesian network can be represented as Influence diagram by augmenting it with utility nodes and decision nodes, where utility node represents the value of a particular event and decision node represents the choices that might be made. Bayesian network has been used in various fields like medicine, forecasting, control, and modeling for human understanding to infer the knowledge using Bayesian Inference [20].

Bayesian inference is the most appealing technique to perform reasoning with Bayesian network, and learning Bayesian network parameters and structures to adapt the changes in the source of information, which can be performed using various methods like variable elimination and arc reversal, conditioning to perform inference in multiply-connected networks, logic sampling, Markov-Chain Monte Carlo methods etc [20].

Different representation techniques have been proposed to represent uncertain data and different uncertain reasoning algorithms have been implemented to get exact answer to the query and to retrieve relevant information from unorganized knowledge facts [16], [17], [19].

Uncertain reasoning employs various types of methodologies to deal with uncertain data to retrieve relevant and unambiguous information from knowledge base, where knowledge base can be represented in various forms according to type of uncertainty and its representation. Major components of knowledge base are concepts, roles, individuals and axioms, where concepts refer facts, roles are relationships between concepts, individuals are instances of concepts which follow properties of concepts, and axioms are relationships among concepts and roles.

Bayesian model and fuzzy logic model are two mathematical models to deal with various types of uncertainty viz. inconsistency, ambiguity, empirical, vagueness, incompleteness and inaccuracy (figure 1) [3]; as reported by Uncertainty Reasoning for the World Wide Web Incubator Group (URW3-XG), where Bayesian model is based on probability theory and fuzzy model is based on fuzzy logic. Probability theory can be used to represent and reason about inconsistency and ambiguity type of uncertainty in knowledge base. Fuzzy model can represent vagueness or fuzziness of facts in knowledge base.

Bayesian model is categorized into Bayesian network and Probabilistic extensions to description logic. Fuzzy logic model is categorized into first order probabilistic approach, fuzzy propositional logic and fuzzy description logic as shown in Figure 2. This paper concentrates mainly on Bayesian model, and available approaches based on Bayesian network for uncertainty representation and reasoning.

Bayesian network is the main part of Bayesian model, which is based on Bayes Theorem and its Product rule, which is main logic behind Bayesian inference. It uses joint probability distribution to represent each node along with its dependency on its predecessors in Directed Acyclic Graph (DAG), where DAG represents the degree of belief of events in the network. Each node will annotate with quantitative probability information i.e. conditional probability table.

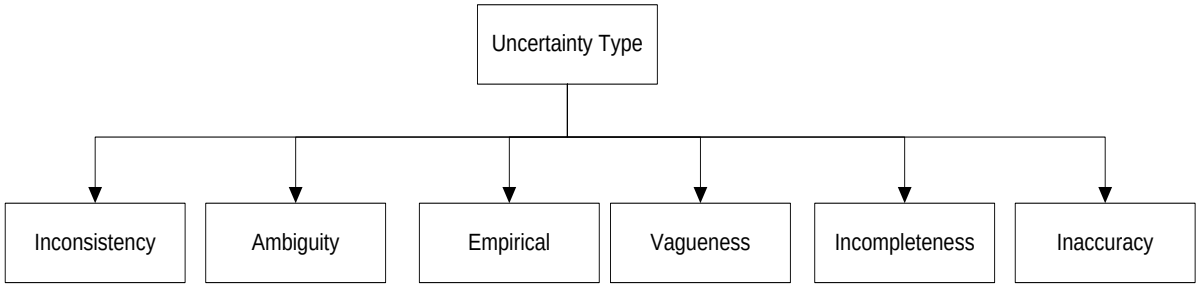


Figure 1. Uncertainty Type [3]

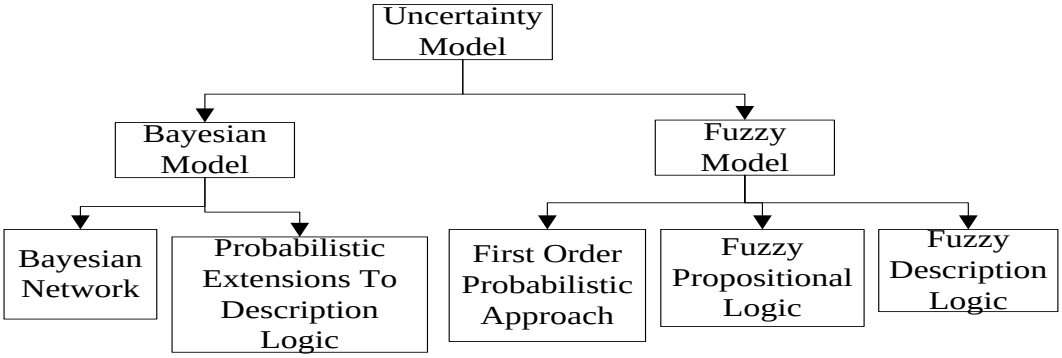


Figure 2. Uncertainty Model [3]

This paper provides some of the major existing representation and reasoning approaches regarding uncertainty representation using Bayesian network. The listed works in this paper can be classified into following two categories

a) Description Logic, Object Oriented and Entity Relation based Model

- P-Classic [22]
- SPOOK [23]
- Probabilistic Relational Model [30]
- DAPER Model [31]

b) Bayesian network with semantic web languages

- Probabilistic extension of RDF [4]
- Extension to RDF(S) [12]
- OntoBayes [25]
- BayesOWL [5]
- PR-OWL 2.0 [9], [11] and[18]

This paper is organized as follows: the paper gives semantics of Bayesian network with an example to make this paper self-contained in the second section. Section 3 presents methodologies based on description logic, object-oriented, and entity relationship. Proposed works based on augmentation of Bayesian network with semantic languages viz. RDF, RDF(S) and OWL are briefly discussed in section 4. A list of works done using PR-OWL and MEBN has been presented in section 5 and conclusion is given in section 6.

2. SEMANTICS OF BAYESIAN NETWORK

Semantics of Bayesian Network (BN) can be viewed in two ways: the first one is the network as representation of joint probability distribution and second one is the encoding of a set of conditional independence statements [1]. The first view as; Bayesian network is a concise specification of any joint probability distribution, where a joint probability distribution will be able to answer a query or calculate the probability of an unknown event with given events. An uncertain attribute, feature, or hypothesis will be represented by random variable as a node in BN. The uncertainty of the dependence is shown as an edge between two nodes in DAG and represented as $P(X_i|parents_i)$ in conditional probability table, where X_i is a node and parentsi is the parent node set of X_i. Each entry in the full joint probability distribution for a node will be calculated by taking values of affecting parent nodes in the form of conjunction of probability of each variable, such as $P(X_1=x_1 \wedge \ldots \wedge X_n=x_n)$ abbreviated as $P(x_1,..., x_n)$.

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$$

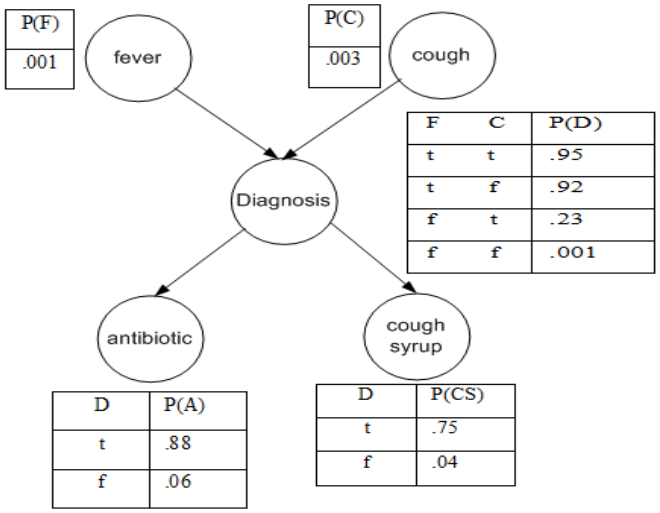


Figure 3 Bayesian Network for Medical Diagnosis

Figure 3 presents an example of Bayesian network, in which, a doctor has to diagnosis a person for fever or cough or both. Doctor can prescribe either antibiotic only or cough syrup only or both according to diagnosis. The probability of correct diagnosis, where person has neither fever nor cough where antibiotic and cough syrup both have been prescribed; will be calculated a-s follows

$P(A \wedge CS \wedge D \wedge F \wedge \neg F \wedge \neg C)$

$= P(A|D)P(CS|D)P(D|\neg F \wedge \neg C)\, P(\neg F)\, P(\neg C)$

$= 0.88 \times 0.75 \times 0.001 \times 0.999 \times 0.997$

$= 0.00065736198$

3. DESCRIPTION LOGIC, OBJECT ORIENTED AND ENTITY RELATION BASED MODEL

Description Logic (DL) [10] can be viewed as stable representation of knowledge base for last decade. To represent uncertain information in description logic some researchers have been used Bayesian network, which can be seen as a part of probabilistic extensions to description logic.

Description logic is a subset of first order logic (FOL) to create a knowledge base. It uses terminological box (T- Box) to create vocabulary of concepts and assertions box (A-Box) to represent assertions about instances of concepts of T-Box or not, which makes easy to reason about knowledge base. It can be represented using resource description framework (RDF), resource description framework schema (RDFS) and web ontology language (OWL). OWL is available in 3 levels: OWL Lite, OWL DL and OWL Full. SHIF (D) is logic behind OWL Lite and SHOIN (D) for OWL DL, where OWL Lite has less expressivity than OWL DL and OWL DL has less expressivity than OWL Full.

3.1 P-CLASSIC

It is one of the first probabilistic versions of the DL Classic [22]. It supports terminological knowledge and uncertainty about properties of individual, the number of fillers for its roles, and the properties of these fillers is represented using Bayesian network. It has proposed inference procedure for probabilistic subsumption: computing the probability that a random individual in class C is also in class D. Knowledge representation formalisms can be based on a) rule-based languages b) object-centered formalism. This approach has integrated probabilities with object-centered language, which allows the specification of a probability distribution over the properties of individuals. Random variables of Bayesian network are the basic properties of individuals, the number of their fillers, and properties of their fillers. Probabilistic subsumption is provided to check the probability of belonging of complex concept C within the set of individuals in D. It gives better degree of overlap between two concepts, which was limited in CLASSIC description logic. It uses p-classes (probabilistic classes) to represent probabilistic component, each of which is a Bayesian network. A set of p-classes will represent probabilistic information of basic properties of individuals and other p-classes for role fillers.

CLASSIC uses two types of statement in its T-Box i.e. concept introductions and concept definitions, but terminology in P-CLASSIC will use only concept definitions and concept introductions are embedded with probabilistic component.

The major drawback of P-CLASSIC is unable to express equality relationship. Since different fillers are disjoint, and the number of fillers for each role is bounded. Some improvements were needed like support of disjunctive concepts, existential quantification, negation on arbitrary concepts (not only primitive ones) and qualified number restrictions. It doesn't include same-as constructor of CLASSIC DL.

3.2 SPOOK: A system for probabilistic object-oriented knowledge representation

This model is compact, modular, natural, and easy to build [23]. It is a unified model, which supports reusability and encapsulation. It can be used to develop and manage ontology of a large complex domain. It can also implement complex

structured domain on the basis of objects and classes where uncertainty is represented using probability.

This is tested on military situation assessment, which have a large number of objects and those are related to one another in various ways. It has four major features like multi-centeredness (i.e. each object can be accessed by a multitude of other objects, in a variety of ways), encapsulation, multi-valued attribute and quantifier attribute, and structural uncertainty. Structural uncertainty is of two type i.e. number uncertainty and reference uncertainty. Number uncertainty implies uncertainty over the number of values of a multi-valued complex attribute and reference uncertainty implies which is uncertainty over the value of a single valued complex attribute.

To perform query, it performs knowledge-based model construction at first and then Bayesian inference algorithm will be used. In the representation of various relationships among objects, it can violate the part-whole property and it cannot treat type uncertainty.

3.3 Probabilistic Relational Model

Probabilistic relational model (PRM) is a structured statistical model, which describes the domain using relational schema augmented with probabilistic distribution that is known as relational logic [30]. Frame-based logical representation is augmented with Bayesian network. Syntax is inspired from frame-based and object-oriented system so, it extends Bayesian network with objects, attributes and relationship.

PRM has three components: relational schema, probabilistic graphical model, and relational skeleton. A relational schema is a logical description of the domain of discourse, which will be transformed into a frame-based representation. It will use the concepts like relation for a class, column for an attribute of a class, and reference slot (opposite is inverse slot) for foreign key. A probabilistic graphical model is depicted using directed acyclic graph and can be represented in a logical formalism. A relational skeleton is specified for each class of a relational schema which will have a set of uninitialized objects.

Random variable will be defined for each uncertain attribute of the object where attributes of a same class or different class may be dependent on one another. To ensure acyclic representation of graph, it uses mainly two graphs: instance dependence graph and class dependence graph.

PRM provides various types of uncertainty like structural uncertainty, attribute uncertainty and class uncertainty, where structural uncertainty is in two forms: reference uncertainty and existence uncertainty. It uses two inference techniques for reasoning and learning purposes viz. exact inference and approximate inference.

3.4 DAPER model

The directed acyclic probabilistic entity-relationship model is a combination of ER-model with directed acyclic graph (DAG) and local distribution classes [31]. ER-model gives pictorial representation of entity classes, relationship classes, attribute classes, and their interconnections. DAG gives dependence among the attribute classes. Local distribution classes are probability distribution for dependence among attribute classes. Therefore, it is the graphical language for probabilistic entity relationship (PER) model and can also implement Plate model, where plate model is similar to PRM but suitable for statistician and PRM is used for computer users. It can perform probabilistic inference about attribute classes.

It can represent restricted relationships, self-relationships, partial relationship existence, and probabilistic relationships. It does not support first order formulas and quantifier, but due to dependence on random variable semantics it can be extended to FOL.

4. BAYESIAN NETWORK WITH SEMANTIC WEB LANGUAGES

This section mainly will go through the proposals based on an extension of RDF, an extension of RDF(S), and two proposals based on extension of OWL i.e. BayesOWL, OntoBayes, and MEBN and PR-OWL 2.0.

4.1 Probabilistic extension of RDF

Fukushige has proposed an implementation to integrate probability with RDF in the form of vocabulary [4]. Framework has been proposed on the basis of RDF and Bayesian network to calculate probability distribution. RDF is widely used language to represent ontology in semantic web and several successful software had been developed to create ontology using RDF and perform reasoning about that. This framework doesn't support any standard query language of semantic web instead Bayesian inference was the sole reasoning technique. The given framework starts by creating a vocabulary of propositions along with probability, which leads to a RDF graph. RDF graph should be converted into Bayesian network to perform reasoning.

An example of probabilistic relationship is borrowed from [6] for metastatic cancer. Vocabulary is created in N3 language [7], which includes propositions, negations, unconditional probabilities, conditional probabilities, observations and posteriors. This model can represent specific classes of problems with less expressiveness.

4.2 Extension to RDF(S)

Holi and Hyvonen have used Bayesian network to compute degrees of overlap or ambiguous uncertainty between concepts of taxonomy [12]. RDF(S) is used to represent concept and partial subsumption. Concepts of taxonomies are represented as different sets in Venn diagram which helps to create overlap table for every concept. It uses a DAG to represent partial overlap where each node represents a concept which is annotated with mass i.e. size of the set. The graph is converted from partial overlap to a solid path structure using breadth first search algorithm. Overlap values will be interpreted as a conditional probabilities and efficient evidence propagation algorithm is used to calculate overlap computation. An individual will be matched with the query concept by computing the degrees of overlap, where degree of

overlap plays significant role to measure the relevance of the concept in information retrieval techniques. BayesOWL can be seen as next methodologies on the basis of this approach.

4.3 BayesOWL

BayesOWL was one of the successful approaches to represent uncertain information based on probabilistic framework, which uses OWL to represent information in knowledge base [5]. It has been elaborated uncertain reasoning and mapping of ontology as two major applications of Bayesian Network.

This framework augments OWL with BN to represent uncertainty and perform uncertain reasoning. It gives a methodology to construct BN using OWL taxonomy for which five structural translation rules have been proposed. It is a mechanism to express OWL ontologies as Bayesian Network by adding additional nodes according to following constructors shown in Table 1.

On the basis of supported constructors, conversion of OWL taxonomy to a BN DAG is performed in two steps a) Structural Translation and b) Conditional Probability Table creation.

For structural translation, it uses two types of nodes a) Concept nodes b) L-nodes. Concept nodes are used for regular concept class. L-nodes are used for modeling relations among concept nodes where relations are logical operations and works as bridge between concept nodes. L-nodes will be created as leaf nodes to avoid cycle. It has two types of properties and can be classified as prior and conditional properties. Figure 4 shows an example of owl:unionOf constructor from [5]

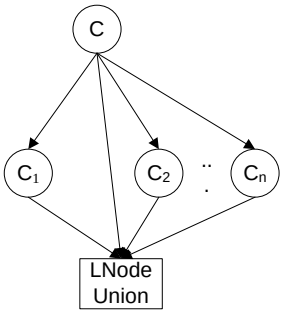


Figure 4 Conversion of union of concept nodes into LNodeUnion for Bayesian Network [5]

Here, Union of concept classes C_i (i=1… n) is defined as C, using constructor "owl:unionOf" as partial BN. Conditional probability table will be created for both types nodes (Concept nodes as well as L-nodes) on the basis of proposed logical relations, which will include prior probability as well as conditional probability.

Piece-wise probability constraints can be used to construct CPTs with help of Decomposed-Iterative Proportional Fitting Procedure (D-IPFP) algorithm. Probabilistic inference plays major role for reasoning purposes in the proposed approach.

Three types of BN connections are allowed: serial connections, diverging connections and converging connections as shown in Figure 5.

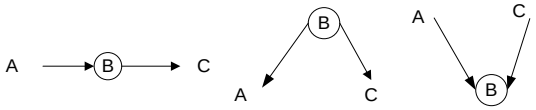


Figure 5 Bayesian Network connections a. Serial b. Diverging c. Converging [5]

BayesOWL supports only three basic reasoning within and across same ontology i.e. concept satisfiability, concept overlapping and concept subsumption, where concept satisfiablility checks whether a perfect description of a concept is available, concept overlapping decides the degree of overlap between a concept and its description, and at last concept subsumption finds whether a concept follows the properties of a given description.

The advantage of BayesOWL is that it reduces the cost and user efforts, since it doesn't enforce to modify OWL and ontologies. In addition to that it is not highly dependent on syntax instead of semantics as well as it can be used to translate either partial ontology or conditional probability table into BN according to user requirement in consistent fashion.

Limitations of BayesOWL: It cannot deal with multi-valued random variable. In translation of ontology, it does not consider the instances, the specific data type, and properties represented by BayesOWL.

4.4 OntoBayes

It is an ontology-driven Bayesian model for uncertain knowledge representation [25] [34]. OntoBayes is a decision-theoretic design analysis. It integrates Bayesian network into web ontology language to annotate the ontology with Bayesian probability along with dependency relationship. OntoBayes deals only with discrete random variable but Boolean random variable as a special case. It does not deal with continuous random variable.

To provide probabilistic integration to OWL, only three classes have been introduced viz. PriorProb, CondProb and FullProbDist. PriorPrab and CondProb have only one property Probvalue and FullProbDist has two properties hasPrior and hasCond. PriorProb deals with unconditional probability and CondProb deals with conditional probability. FullProbDist deals with joint probability distribution table.

It uses two different graphical representation one for Bayesian graph and other for OWL. Bayesian graph depends on properties of OWL classes, and OWL graph depends on classes and properties. Although both graphs will have subject, predicate and object in the graph, but Bayesian graph will use only one predicate <rdfs:dependsOn>.

One major advantage of OntoBayes is that it can represent cyclic dependency and can deal with multi-valued random variable. For more expressivity, some approaches based on

objects, entity relationship, and first-order logic is proposed, which are discussed in next section.

Table 1. Supported Constructor [5]

Constructor	DL Notation	Class Axiom	Logical Operator
rdfs:subClassOf	C1 ⊑ C2	*	
owl:equivalentClass	C1 ≡ C2	*	
owl:disjointWith	C1 ⊑ ¬C2	*	
owl:unionOf	C1 ⊔ … ⊔ C2		*
owl:intersectionOf	C1 ⊓ … ⊓ C2		*
owl:complementOf	¬C		*

4.5 MEBN and PR-OWL 2.0

Probabilistic Web Ontology Language (PR-OWL) [11] provides uncertainty representation using OWL constructs based on Multi-Entity Bayesian Network (MEBN) logic [9] and it is successfully implemented in UnBBayes [21] graphical user interface to model a probabilistic ontology [2], [8], [13], [14], [15], [18], [24], [26], [27] and [37]. MEBN logic is an extended form of Bayesian network along with expressive power of first order logic formula, where Bayesian network will encode a set of dependent evidences in graphical form and degree of belief of evidence can be interpreted with the help of constraints specified using first-order formulas. Basic constructs of MEBN are MTheory, MFrag, random variable, and entities.

MTheory is the label to combine multiple MFrags together and a probabilistic ontology must have at least one MTheory to represent the model. It will ensure the consistency of probabilistic ontology by checking unique joint probability distribution table. To perform the query, one has to provide domain specific entity instances i.e. knowledge base then Bayesian inference will be used for query and learning of new evidences.

MFrag is the combination of random variables, and fragment graph, where random variable will be represented as resident node along with conditional probability distribution table and fragment graph will show the dependence among random variables in the form of directed acyclic graph. A MFrag will give a template of a Bayesian network, which can be instantiated multiple times by binding its arguments to domain entity identifiers to create instances of its random variables. Resident node will work as input node to the other MFrag of ontology, where union of input node and resident node will be used as parent nodes in the fragment graph that is why the probability distribution table of the child node will have entries for the states of all its parent nodes. Here, mutual exclusive and exhaustive number of states of random variable will be represented using entities.

The advantage of MEBN over simple Bayesian network is that it can represent repeated structure of Bayesian network and it allows dynamic infinite number of instances of a Bayesian network structure. MEBN logic can represent various types of uncertainty like attribute value uncertainty, existence uncertainty, number uncertainty, referential uncertainty, structural uncertainty, and type uncertainty [9].

MEBN logic is successfully implemented using PR-OWL upper ontology, which is based on basic model as shown in figure 5. Here, ovals represent class and arc represents relationship between classes.

PR-OWL 1.0 has two shortcomings 1) it cannot provide mapping to properties of OWL and 2) although it provides concept of meta-entities for the definition of complex types, it does not have type compatibility with OWL. Therefore, PR-OWL 2.0 has been proposed to overcome these shortcomings to provide better modeling of domain information [18].

5. PROPOSED MODELS USING PR-OWL FRAMEWORK

Although PR-OWL is in initial phase of establishment and it has not been widely accepted by semantic web community [36], but some relevant proposed works discussed in this section will present importance and applicability of it.

5.1 Service based situation awareness

There are 3 levels of situation awareness presented in [34], 1) Perception, 2) Understanding, and 3) Projection, understanding type of situation awareness have been considered in [28] for service-oriented information technology environments. For situation assessment of status or health of IT services understanding (the second level) is more suitable than perception (the first level), because perception is more probable for inaccuracy and incomplete.

It has been proposed to develop a PR-OWL framework based ontology definition that supports automated reasoning with uncertain service-based situation awareness.

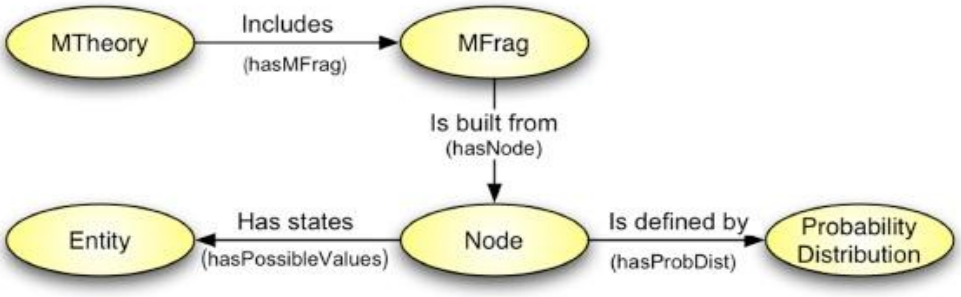


Figure 6 PR-OWL basic model [13]

5.2 Research on Interactive Behaviour Analyzing in New-type Distributed Software System

It has been tried to extract interactive behaviours produced by loosely-coupled software entities [33]. Behaviour of software entities is investigated on the basis of historical knowledge and current practical evidences. It has adopted the case based reasoning along with case reasoning.

It has been considered the new-type distributed software to supervise its group behaviour regarding behaviour of dynamic, accidental, correlative, and repeating properties. It has used MEBN to represent group behaviour as well as knowledge fusion of monitored evidence information and experimental knowledge. Fusion of problem space and solution space is useful to accurately analyze software behaviour credibility for current context.

5.3 PROGONOS: Probabilistic Ontology for Net-centric operation systems

It is Predictive Situation AWareness (PSAW) model to be used for U. S. Navy's FORCENet for prediction about battlefield, where huge amount of data is collected from sensors, human intelligence, and others [29]. It is modeled using PR-OWL framework, where it is based on two modules, the first module is used for performing reasoning and second module is used for simulation.

5.4 Probabilistic Risk Assessment for Security Requirements: A Preliminary Study

Risk assessment is the process of identifying risk factors and relationship among risk factors, which includes judgement as well as meta-judgement about the degree of certainty that they have in their judgements [35]. To represent the degree of certainty, it has been proposed a model, where security requirements and their causal relationships are represented using MEBN logic.

5.5 A Model of an Ontology Oriented Threat Detection System (OOTDS)

Object-oriented threat detection system is dynamic behaviour based system, which will be automatically upgraded according to changes in environment [14]. It is mainly based on two ontologies a) Threat detection ontology and b) Threat detection learning ontology, where the first one is used for threat detection and the second one is used for learning the dynamic changes in environment. Here, ontologies are created using PR-OWL model again.

6. CONCLUSIONS

Bayesian network can play a pivotal role in the representation of uncertainty and performing reasoning in semantic web. Survey on the Bayesian network for the representation of uncertainty has been discussed on the bases of description logic, object oriented, entity relational model approaches, as well as core works done so far for augmentation of Bayesian network with semantic web languages viz. RDF, RDF(S), and OWL to represent uncertainty in knowledge bases, which provides efficient ways to perform reasoning to extract relevant information for specific purposes. The list of works done using PR-OWL has been discussed, which is a domain ontology based on MEBN logic and its relevance can been experienced by applying in various applications of real life.

7. ACKNOWLEDGEMENTS

We are very thankful to Dr. Rommel N. Carvalho and Dr. Paulo Cesar G. da Costa for giving permissions to use some contents and figures from their Ph. D. theses and papers.

8. REFERENCES

[1] Stuart J. Russell & Peter Norvig (2003) Artificial Intelligence – A Modern Approach, Second Edition, Pearson Education, Inc.

[2] Kumar Ravi & G. Aghila (2013) Probabilistic Uncertainty Representation in Semantic Web, International Journal of Information and Computation Technology, Volume 3, Number 1 (2013), [SPECIAL ISSUE] pp. 27-33.

[3] Kenneth J. Laskey and K. B. Laskey, (2008) "Uncertainty reasoning for the world wide web: Report on the URW3-XG incubator group," W3C, URW3-XG, 2008. http://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/

[4] Yoshio Fukushige (2005) Representing Probabilistic Knowledge in the Semantic Web, available http://www.w3.org/2004/09/13-Yoshio/PositionPaper.html

[5] Z. Ding (2005) BayesOWL: A Probabilistic Framework for Semantic Web. Doctoral dissertation. Computer Science and Electrical Engineering. 2005, University of

Maryland, Baltimore County: Baltimore, MD, USA. p. 168

[6] J. Pearl (1998) Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 1988.

[7] An RDF language for the Semantic Web Available: http://www.w3.org/DesignIssues/Notation3.html

[8] Rommel Carvalho, Kathryn Laskey, Paulo Costa, Marcelo Ladeira, Laecio Santos and Shou Matsumoto (2010) UnBBayes: Modeling Uncertainty for Plausible Reasoning in the Semantic Web, Semantic Web, Gang Wu (Ed.), ISBN: 978-953-7619-54-1, InTech, Available from: http://www.intechopen.com/books/semanticweb/unbbayes-modeling-uncertainty-for-plausible-reasoning-in-the-semantic-web

[9] K. B. Laskey (2008) MEBN: A language for first-order Bayesian knowledge bases, Artificial Intelligence 172 (2008) 140–178

[10] F. Baader & W. Nutt (2003) Basic Description Logics. Chapter in The Description Logics Handbook: Theory, Implementation and Applications. Baader, F.; Calvanese, D.; McGuiness, D.; Nardi, D.; Patel-Schneider, P.; editors. 1st edition, [47-100]. Cambridge, UK: Cambridge University Press

[11] P. C. G. Costa (2005) Bayesian Semantics for the Semantic Web, Ph. D. thesis; George Mason University.

[12] M. Holi, E. Hyvonen (2004) A method for modeling uncertainty in Semantic Web taxonomies, in: Proceedings WWW-2004, ACM Press, 2004, pp. 296–297

[13] Rommel N. Carvalho, Kathryn B. Laskey, Paulo C. G. Costa, Marcelo Ladeira, Laécio L. Santos, and Shou Matsumoto (2013) Probabilistic Ontology and Knowledge Fusion for Procurement Fraud Detection in Brazil, F. Bobillo et al. (Eds.): URSW 2008-2010/UniDL 2010, LNAI 7123, pp. 19–40, 2013. Springer-Verlag Berlin Heidelberg.

[14] Karol Banczyk, Henryk Krawczyk (2008) A model of an Ontology Oriented Threat Detection System (OOTDS), Proceedings of the 2008 1st International Conference on Information Technology, IT 2008, 19-21 May 2008, Gdansk, Poland

[15] Daniel Vincen, Dafni Stampouli, Gavin Powell (2009) Foundations for System Implementation for a Centralised Intelligence Fusion Framework for Emergency Services, 12th International Conference on Information Fusion Seattle, WA, USA, July 6-9, 2009

[16] S. Tobies (2001) Complexity Results and Practical Algorithms for Logics in Knowledge Representation, PhD thesis, RWTH Aachen, Germany.

[17] I. Horrocks, U. Sattler (2005) A tableaux decision procedure for SHOIQ, in: Proceedings IJCAI-2005, 2005, pp. 448–453 (Extended version to appear in J. Autom. Reason.).

[18] Rommel N. Carvalho, Kathryn B. Laskey, and Paulo C.G. Costa (2013) PR-OWL 2.0 – Bridging the Gap to OWL Semantics, F. Bobillo et al. (Eds.): URSW 2008-2010/UniDL 2010, LNAI 7123, pp. 1-18, 2013. Springer-Verlag Berlin Heidelberg.

[19] T. Lukasiewicz (2002) Probabilistic default reasoning with conditional constraints, Ann. Math. Artif. Intell. 34 (1-3)35-88.

[20] Ronan Daly, Qiang Shen and Stuart Aitken, (2011) Learning Bayesian networks: approaches and issues, The Knowledge Engineering Review, Vol. 26:2, 99-157. Cambridge University Press, 2011, doi: 10.1017/S0269888910000251

[21] Paulo C. G. COSTA, Kathryn B. Laskey, PR-OWL: A Framework for Probabilistic Ontologies "UnBBayes - the UnBBayes site," http://unbbayes.sourceforge.net/

[22] D. Koller, A. Y. Levy, and A. Pfeffer, (1997) P-CLASSIC: A Tractable Probabilistic Description Logic. Paper presented at the Fourteenth National Conference on Artificial Intelligence (AAAI-97), July 27-31. Providence, RI, USA

[23] Pfeffer, A.; Koller, D.; Milch, B.; and Takusagawa, K. T. (1999) SPOOK: A System for Probabilistic Object-Oriented Knowledge Representation. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 541-550, July 30 – August 1. Stockholm, Sweden Sowa, J. F. (2000). Knowledge Representation: Logical, Philosophical, and Computational Foundations. Pacific Grove, CA, USA: Brooks/Cole.

[24] R. N. Carvalho (2008) Plausible reasoning in the semantic web using Multi-Entity bayesian networks - MEBN, M.Sc., University of Braslia, Brasilia, Brazil, Feb. 2008.

[25] Y. Yi (2007) A framework for decision support systems adapted to uncertain knowledge. Ph.D. thesis; Fakultat fur Informatik der Universitat Fridericiana zu Karlsruhe.

[26] Rommel N. Carvalho (2011) Probabilistic Ontology: Representation and Modelling Methodology, Ph. D. thesis; George Mason University, Brazil.

[27] Rommel N. Carvalho, Laécio L. Santos, Marcelo Ladeira, Paulo C. G. Costa (2007) A GUI Tool for Plausible Reasoning in the Semantic Web using MEBN. In Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007). Mourele, L.; Nedjah, N.; Kacprzyk, J.; and Abraham, A. (eds.); pp. 381-386. October 22-24, 2007, Rio de Janeiro, Brazil.

[28] Stephen. C. Dinkel, William Hafner, Paulo Costa, Sumitra Mukherjee (2011) Uncertainty Reasoning for Service-based Situational Awareness Information on the Semantic Web, 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Miami Beach, FL

[29] R. N. Carvalho, P. C. G. Costa, K. B. Laskey, and K. Chang (2010) "PROGNOS: predictive situational awareness with probabilistic ontologies," in Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, Jul. 2010.

[30] Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer and Ben Taskar, Probabilistic Relational Models

[31] David Heckerman, Christopher Meek, and Daphne Koller (2004) Probabilistic Models for Relational Data, March 2004, Microsoft Research and Stanford University, Technical Report, MSR-TR-2004-30

[32] R. N. Carvalho, R. Haberlin, P. C. G. Costa, K. B. Laskey, and K. Chang (2011) "Modeling a probabilistic ontology for maritime domain awareness," in Proceedings of the 14th International Conference on Information Fusion, Chicago, USA, Jul. 2011.

[33] Changyun Li, Junfeng Man, Zhibing Wang, and Xiangbing Wen (2010) Research on Interactive Behavior Analyzing in New-type Distributed Software System, Proceedings of 2010 Conference on Dependable Computing (CDC'2010) November 20-22, 2010, Yichang, China

[34] M. R. Endsley (1995) Toward a Theory of Situation Awareness in Dynamic Systems, Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 37, pp. 32-64, 1995.

[35] Seok-Won Lee (2011) Probabilistic Risk Assessment for Security Requirements: A Preliminary Study, 2011 Fifth International Conference on Secure Software Integration and Reliability Improvement, IEEE computer society, DOI 10.1109/SSIRI.2011.12

[36] Amandine Bellenger, Sylvain Gatepaille (2008) Uncertainty in Ontologies: Dempster-Shafer Theory for Data Fusion Applications, 2008.

[37] Kumar Ravi, Sheopujan Singh (2013) Risk Prediction for Production of an Enterprise, International Journal of Computer Applications Technology and Research (IJCATR), Volume 2 Issue 3 May-June 2013, doi: 10.7753/IJCATR0203.1006

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 539 - 547, 2013, ISSN: 2319–8656

Strategies in the Design of Low Power Wireless Sensor Network for the Measurement and Monitoring of Physiological Parameters

D.Vishnu Vardhan
Department of ECE
JNTUA College of Engineering
Pulivendula-516390-India

Y.Narasimha Murthy.
Department of Electronics
Sri SaiBaba National College
Anantapuramu-515001-India

M.N.Giri Prasad
Department of ECE
JNTUA Collegeof Engineering
Anantapuramu-515001-India

Abstract: This research paper presents the strategies involved in the design of a low power wireless sensor network for the measurement and monitoring of physiological parameters like body temperature, pulse rate and respiration rate of a patient .Ultra low power MSP430F1611 microcontroller from Texas instruments (TI) and the low power transceiver CC2500 ZigBee module from chicon are used in the present design. The TMP100 a digital temperature sensor is used for the measurement of body temperature ,the Free Scale MMA7260Q tri-axial accelerometer is used to measure the respiration rate and a simple pulse rate module is designed using the IR LEDs for the measurement of heart rate. A GUI is developed based on Visual Basic software to display the results. Also the results are displayed on the 2x16 LCD module . In addition , a website with URL www.vishnuhealth.com is designed to display the results .To achieve the low power objective ,an attempt is made to apply the Dynamic voltage and frequency scaling algorithm in addition to all the possible hardware precautions and MAC protocols.

Keywords: Low power wireless sensor, MSP430 microcontroller,Zigbee module ,Accelerometer sensor, Temperature sensor, MAC Protocol.

1. INTRODUCTION

Battery operated wireless sensor embedded system design is the order of the day.But the major limitation for this type of designs is the power consumption which affects the battery life.So,the researchers all over the world are seriously working in this field of low power design to find suitable solutions. The low power architectures, protocols, topologies are the prime concern in the design flow of low-power wireless sensor networks. Especially in the case of battery operated embedded systems which are exclusively designed for bio-medical applications, this fact has utmost importance.

In fact it is not enough if one considers the power consumption for the hardware sub-systems in order to optimize the wireless sensor network, but must consider the power consumption effects of algorithms, network protocols etc[1] i.e it is very essential to balance the power optimized architectures with network algorithms. Hence while designing a low power wireless sensor system for monitoring the bio-medical parameters ,both the software as well as hardware aspects must be equally taken care of.

The transceiver and the microcontroller are the two important candidates in consuming power in a wireless design . So, it is very essential to limit the power consumption of the processor/microcontroller to enhance the battery life. New microcontroller architectures focused in low power systems try to solve this problem by providing hardware support to adapt power consumption to application performance requirements. For example the strong ARM microprocessor from Intel consumes around 400mW of power during the execution of the instruction of the instructions while the ATmega 103L AVR microcontroller from Atmel consumes about 17mW[2,3] .But this low power consumption also affects the performance of the processor drastically. Next, Texas Instruments has released MSP430 family of microcontrollers with five different low power operational modes .They consume around 1.2mW in fully operational mode and 0.3µW in the deepest sleep

mode[4].But this change between the operating modes is also a power consuming task ,especially in applications where fast data sampling and data acquisition ,it is not a preferred option.The current consumption of a MSP430 microcontroller in active mode and in other low power modes is shown in Figure 1.below.

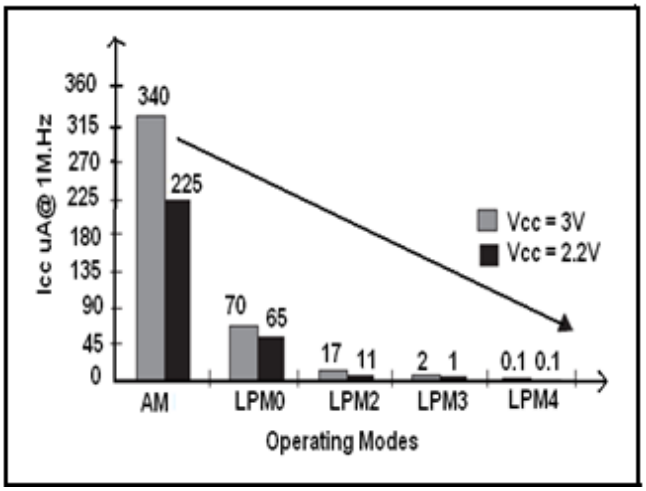


Figure 1. Graph showing the current consumption in different operating modes of MSP430 family.

The power consumed by the CMOS chip is proportional to the frequency and also proportional to the square of the supply voltage as shown below.

$$P \alpha f V_{DD} \quad (1)$$

So, it can be observed that ,reducing the supply voltage from 5.0V to 3.3V reduces the power by nearly 55%.Another effect is reducing the clock frequency .The reduced voltage or voltage scaling at the lowest clock frequency of 60M.Hz and with a V_{DD} of 0.9 V is found to consume around 1/5 th of the energy per instruction that is required at peak performance. i.e by properly decreasing the supply voltage and frequency one can reduce the power dissipation in a design. This is the basic idea behind the Dynamic Voltage and Frequency scaling(DVFS)[5].

Coming to the wireless sensor networks, to establish the wireless communication, both a transmitter and a receiver are required in a sensor node. The basic task here is to convert a bit stream coming from a processor or microcontroller and convert them to and from radio waves [6].Now a days a single device is used that combines both these the two tasks in a single entity. Such combined devices are called transceivers. But these transceivers usually consume large power of around 70% of the total power. There are several factors that affect the power consumption characteristics of a transceiver, including the Type of modulation scheme used, data rate, transmit power (determined by the transmission distance), and the operational duty cycle [7].But most of these transceivers can be configured by the user to set the power level by setting them in various distinct modes of operation like Transmit, Receive, Idle, and Sleep modes.An important fact found in most of the transceivers is that, operating in idle mode results in significantly high power consumption, almost equal to the power consumed in the Receive mode. Thus, it is important to completely shut down the transceiver rather than transitioning to idle mode, when it is not transmitting or receiving any data [8]. Some transceivers can support more diverse power levels, for example the CC2500 has Low current consumption (13.3 mA in RX,250 kbps, input 30 dB above sensitivity limit)[9].

A very important feature of followed in wireless sensor network to reduce power consumption, is that the nodes remain sleeping until they need to undertake a specific task. At some defined time, a sensor node will wake up and perform a measurement. An external event also can trigger this wake-up.The node can then decide to communicate the gathered information to a neighbor and send it a message. Unfortunately, the neighbor might be sleeping to save energy. The node must thus keep sending the information until the neighbor awakens and acknowledges receipt of the information. If a node needs information from a neighbor, it can transmit a request until it receives a response. Alternatively, the requesting node can stay awake and wait until the neighbor decides to send the information spontaneously.Also ,the nodes share a single medium for communication and the performance of the network mainly depends on the effective sharing of this medium by the nodes.The MAC protocol controls the communication nodes and streamlines the sharing of the common wireless medium by the nodes. But the limitations of the MAC protocols are packet collision ,idle listening ,overhearing and overhead[10].To overcome the energy consumption and to implement energy efficient MAC protocol the clustered topology techniques are being followed[11].In the present design a fixed TDMA method at the MAC sub layer is adopted. This TDMA provides the advantage of sending the non active nodes into sleep mode. This leads to low power consumption automatically and the battery life is extended[12].

2. HARDWARE DETAILS

The hardware system consists of three sensors , a temperature sensor TMP100 from TI , a pulse rate sensor and an accelerometer based respiration sensor. All the three sensors are interfaced to the MSP430F1611 microcontroller by using suitable circuitry. A 2x16 LCD module is interfaced to the microcontroller for the display of the data and a GUI , based on Visual Basic software is developed to display the measured parameters. A website with URL www.vishnuhealth.com is also designed to display the results. The block diagram of the design is shown in Figure 2 and the photograph showing the hardware arrangement is shown in Fig.3.

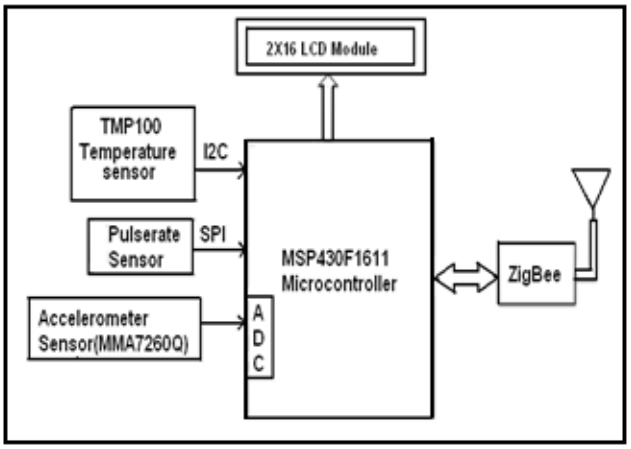


Figure 2. Block diagram of the design

2.1.MSP430 Processor

MSP430F1611 is a 16 bit microcontroller with Ultra low-Power Consumption of 280 µA at 1 MHz, 2.2 V in active mode and 1.6µA in standby mode and 0.1µA in off mode.The supply voltage range is around 1.8 V . . . 3.6 V. and available in 64-Pin Quad Flat Pack (QFP). The architecture of this controller is combined with five low power modes to achieve extended battery life in portable measurement applications. This device features a powerful 16-bit RISC CPU, 16-bit registers, and constant generators that attribute to maximum code efficiency. The digitally controlled oscillator (DCO) allows wake-up from low-power modes to active mode in less than 6 µs. It is configured with two built-in 16-bit timers, a fast 12-bit A/D converter, dual 12-bit D/A converter, one or two universal serial synchronous/asynchronous communication interfaces (USART), I2C, DMA, and 48 I/O pins. In addition, the MSP430F161x series offers extended RAM addressing for memory-intensive applications and large C-stack requirements. Typical applications include sensor systems, industrial control applications, hand-held meters, etc[13].

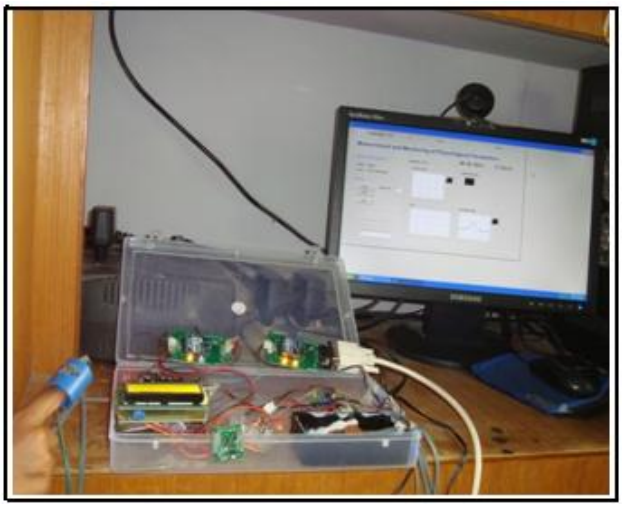


Figure 3. Photograph showing the hardware design

2.2.Temperature Measurement

The temperature of the body is measured by using a simple digital temperature sensor TMP100 which supports I2C standard. Actually,this sensor is mounted within the wrist strap ,positioned in such way that it is in contact with the skin,allowing it to measure the external temperature of the skin. From the skin temperature ,the body temperature is estimated. There can be different methods to estimate the exact body temperature from the skin temperature ,but it is observed that the body temperature is nearly 5 ~ 5.2^{o}C higher than the skin temperature when the body temperature is measured at the ear by a standard thermometer by a medical practitioner.[14]

The TMP100 is a digital temperature sensor which can be operated over a temperature range of -55^{0} C to +125^{0}C . This temperature sensor does not require any complicated signal conditioning circuitry. This sensor is interfaced to MSP 430microcontroller by using the port pins P2.0 and P2.1 The interest in using this temperature sensor is , when operated in one-shot mode, the TMP100 goes into shut down mode after each conversion is completed and consumes a typical current of 0.1uA between conversions[15]. This will justify the low power system design .This temperature sensor is used for the

measurement of body temperature at neck ,wrist and at upper arm as suggested in [16].

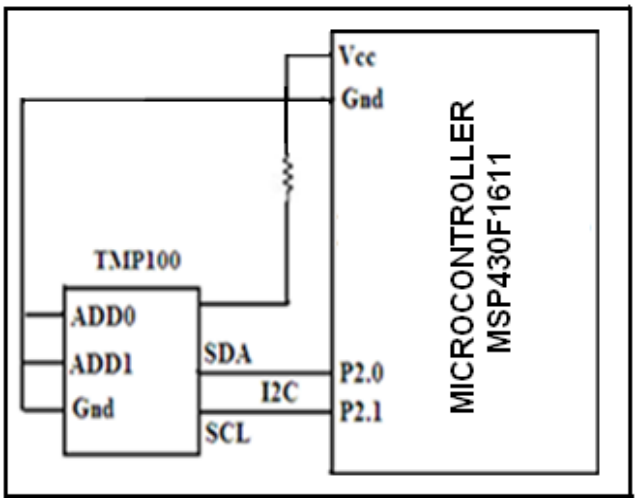


Figure 4. Interfacing of Digital temperature sensor TMP100

The interfacing of the temperature sensor TMP100with the ports of MSP430 microcontroller is shown in Figure 4.

2.3. Pulse rate measurement sensor

Pulse rate denotes the number of heartbeats per second and is usually expressed in beats per minute (bpm). In adults, a normal pulse rate is around 60 to 100 times a minute during resting condition. The resting pulse rate is directly related to the health and fitness of a person and hence is a very crucial physiological parameter. It can be measured at any spot on the body where the pulse is felt with fingers. The most common places are wrist and neck. From this heart rate in bpm is evaluated easily.

To measure the pulse rate a simple low cost pulse-oxymeter is constructed using the arrangement as shown in the block diagram in Figure 5. The main principle of working is based on near infrared spectroscopy using the light of wave length 700-900nm.At these wavelengths most tissues do not absorb light other than hemoglobin.

Basically, the device consists of an infrared transmitter LED and an infrared sensor photo-transistor. The transmitter-sensor pair is clipped on one of the fingers of the patient. The LED emits infrared light to the finger of the subject. The photo-transistor detects this light beam and measures the change of blood volume through the finger artery. The changing blood volume with heartbeat results in a train of pulses at the output of the photo diode, the magnitude of which is too small to be detected directly by a microcontroller. So, a two stage high gain, active low pass filter is designed using two operational amplifiers (Op-Amps) to filter and amplify the signal to appropriate voltage level so that the pulses can be counted by the microcontroller[17].

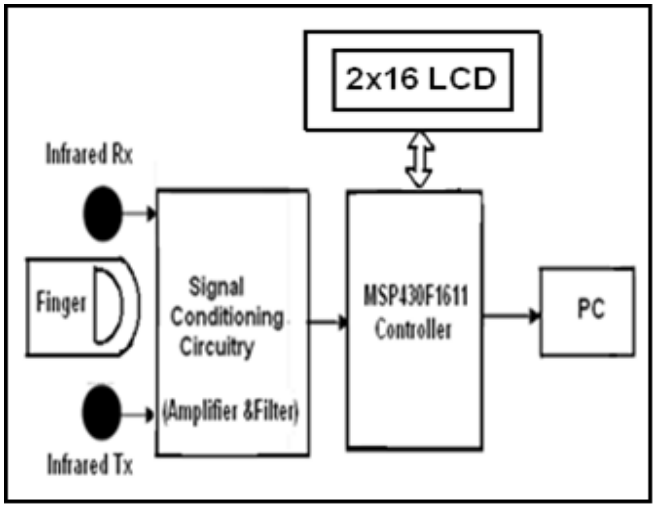


Figure 5. Block diagram of Pulse rate measurement

2.4. Respiration Measurement

Human respiration rate is measured when a person is at rest and involves counting the number of breaths for one minute by counting how many times the chest rises. An optical breath rate sensor can be used for monitoring patients during a magnetic resonance imaging scan. Respiration rates may increase with fever, illness, or other medical conditions. So, the measurement of human respiration is always an important physiological parameter in medical diagnosis.The small rotations at the chest wall due to breathing or during continuous speech provide a valuable information to the doctor . In recent times, the research in the area of accelerometer based respiration techniques gained momentum. This idea as first implemented by Bates et al[18].

In the present design the respiration rate is measured using the Free Scale MMA7260Q tri-axial accelerometer chip .The MMA7260Q is a low cost capacitive micro machined accelerometer features signal conditioning, a 1-pole low pass filter, temperature compensation and g-Select which allows for the selection among 4sensitivities. Zero-g offset full scale span and filter cut-off are factory set and require no external devices. Includes a Sleep Mode that makes it ideal for handheld battery powered electronics [19]. This provides a good solution for XY

and XYZ tilt sensing with a sensitivity of 800mV/g in 3.3V applications. All of these accelerometers will experience acceleration in the range of +1g to -1g as the device is tilted from -90 degrees to +90 degrees. A tri-axial accelerometer is a device that measures the acceleration in three orthogonal directions (sensing axes). An accelerometer can be used to sense vibrations (e.g. the vibration of a machine), orientation (e.g. in human activities monitoring) and hence the respiration rate of human beings .

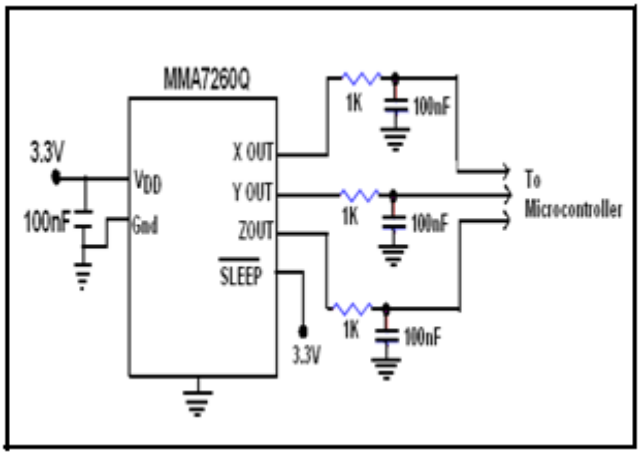


Figure 6.Accelerometer Circuit for respiration measurement

The interfacing circuit of accelerometer sensor MMA7260Q to the microcontroller is shown in Fig.6.The respiration rate is displayed on the 2x16 LCD module. The average resting respiratory rate of an adult is around 24 breaths per minute and of a child is around 18 breaths per minute [20] .The present system is used to measure the respiration rate of adults and children after wrapping the sensor around the chest .The results are found to be satisfactory.

2.5. ZigBee Communication

ZigBee is a low-power, low-cost, wireless networking standard. This low power allows longer life for smaller batteries and provide low cost solutions to many wireless sensor applications. ZigBee operates in the industrial, scientific and medical radio bands (ISM)with 868 MHz, 915 MHz, and 2.4 GHz in different countries. The technology is intended to be simpler and less expensive than other WPANs such as WiFi, Bluetooth and ZigBee . Of these, ZigBee is the most promising standard owing to its low power consumption and simple networking configuration.[v]. Network devices, whether wired or wireless, are commonly described by the Open Systems Interconnection (OSI) reference model. This abstraction model was developed bythe International Standards Organization (ISO), starting in the 1980 description of communication-related protocols and services. The generic seven-layer model is applied to all net work and media types.[vi].In the present work the Zigbee module CC2500 from TI is used. The CC2500 is a low cost true single chip 2.4 GHz transceiver designed for very low power wireless applications. The circuit is intended for the ISM (Industrial, Scientific and Medical)and SRD (Short Range Device) frequency band at 2400-2483.5 MHz.This ZigBee provides the low power features of 400 nA SLEEP mode current consumption, Fast startup time: 240 us from SLEEP to RX or TX mode(measured on EM design) , Wake-on-radio functionality for automatic low-power RX polling and Separate 64-byte RX and TX data FIFOs (enables burst mode data transmission)[21].CC2500 is configured using the SmartRF® Studio software, available for download from http://www.ti.com.

3. SOFTWARE DETAILS

The software development environment of the proposed system is a cross compiler IAR Embedded Workbench which is designed for MSP430 micro-controller by IAR Company[22]. The IAR Workbench IDE is a very powerful Integrated Development Environment that allows to develop and manage complete embedded application projects .It's efficient complier performance and ability to support multiple development tools are its positive points. In System Programming is programming or reprogramming the on-chip flash memory, using the boot-loader software and a serial port. The MSP430 Microcontroller provides on-chip boot-loader software that allows programming of the internal flash memory over the serial channel. Philips provides a utility program for In-System programming called Flash magic Software [23].Visual basic another very useful software which can be readily used to design Graphical User Interface(GUI) to display the measured parameters both in numeric form or graphical forms. This will provide for a good analysis of the results.

4. RESULTS & CONCLUSIONS

The low power wireless sensor network for the measurement of biomedical parameters is designed .The measurement are made over a period of time continuously and the results are displayed on the 2x16 LCD module and also on the Graphical User Interface(GUI) developed using the Visual Basic softwrare. The same results are also made available on the website designed for this purpose with the URL http:// www.vishnuhealth.com.The measurements are made on different persons both male and female at different intervals of time and compared the results with standard clinical instruments. The present readings are found to be satisfactory with in an error of 5%.Since the basic objective of our work is the design of low power ,low cost wireless sensor design, this error is not be taken seriously into account .The body temperature is around 38^0C ,heart beat is 72bpm and the respiration rate is 27 for adults and these values are displayed on the GUI shown in Figure 7.

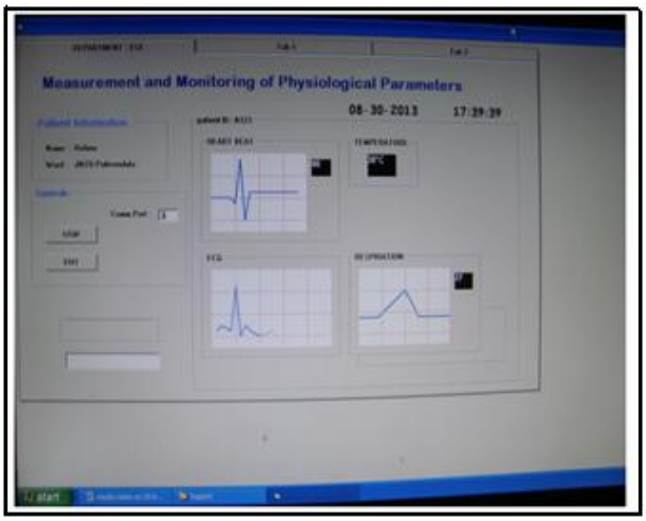


Figure 7. Graphical User Interface (GUI) showing the results

To measure the current consumption, algorithm execution time and the clock frequency ,the methodology suggested by Cebrian[24] is adopted. Though there are certain limitations in this method ,due to its simplicity in implementation ,the authors adopted this methodology. The block diagram used for this implementation is shown in Figure 8.The microcontroller current consumption is obtained indirectly by measuring the voltage drop across a shunt resistor. A precision digital multimeter is used to measure the small currents of the order of few µA at K.Hz clock frequencies to mA at M.Hz clock frequencies. Proper filtering must be done to avoid the noise.

The algorithm execution time is also equally important for low power design.So,this algorithm execution time is measured using the PC DAQ card. The clock frequency of the MSP430 microcontroller is measured using the digital frequency meter.

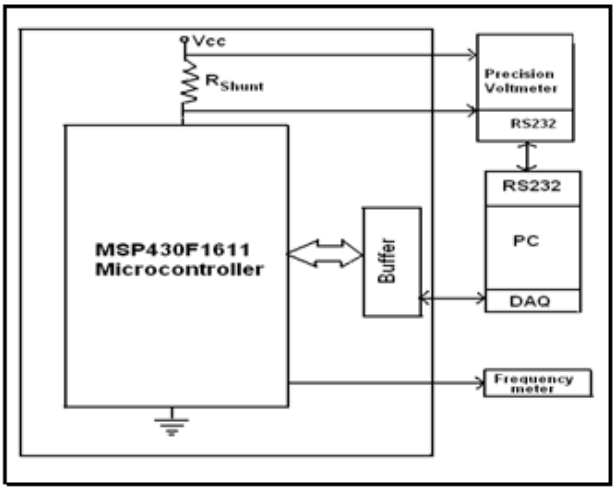


Figure 8. Measurement of current consumption

The power consumed by each subcomponent during the execution is evaluated using the relation

$$\text{Power } P = (I_{mean}.V_{DD}.\tau)/T \quad (2)$$

where V_{DD} is the supply voltage , τ is the effective computation time of the algorithm ,I_{mean} is the average current calculated and T is the maximum time interval of the algorithm execution.

As mentioned in our earlier discussion ,in the total power budget there are two important candidates one is the microcontroller and the other is the transceiver. Between the two , the transceiver consumes around 70% of the total power. By the application of suitable TDMA based clustered MAC protocols the power consumption is brought down considerably to around 62%. Power consumption in a microcontroller during the detection of the signal is more than the power consumed during the sampling of the signal .The variation of power and energy consumption in a MSP430 controller with the supply voltage is shown in the Figure 9.

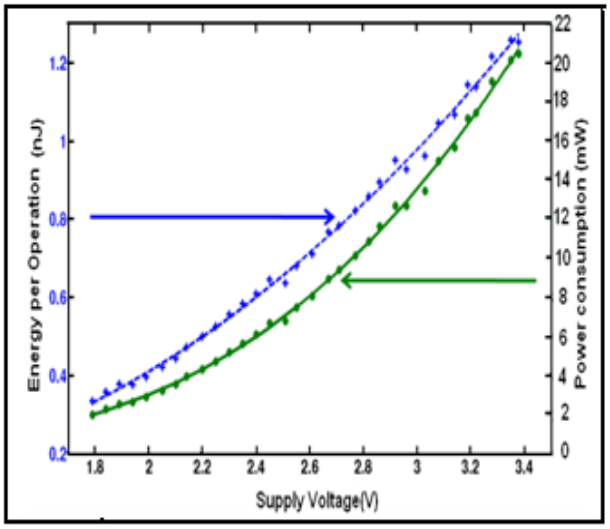


Figure 9. Energy and Power consumption with supply voltage in a MSP430 microcontroller

It is very clear that at higher voltages the power dissipation is more. It is also observed that the power dissipation during the data processing is more than during the transmission of data.So,it is evident that the higher supply voltages lead to larger power dissipation. Hence the dynamic voltage and frequency protocol is applied by operating the microcontroller in LPM3 mode, where the supply voltage levels are decreased during the data transmission automatically using proper algorithm . But, this is achieved at the cost of increased execution time. Similarly the frequency of the oscillator is also suitably modified during the data sensing and transmission as well as during data processing.The total power consumption in the present design is found to be 104mW.

The TMP 100 temperature sensor consumed a typical current of 0.3uA and a power consumption of around 1mW between two conversions in active mode .But when the display is also considered it is about 2.78 uA .In the present design the power consumption of TMP100 sensor is 2.2% of the total power. The commercial pulse oxymeters consume a power of 20 ~ 60mW.In this the LEDs consume a bulk of the total power. The present low-cost pulse measuring device consumed a power of about 6.76 mW. This is about 6.5% of total power consumption. This may not be the lowest power consumption as some body has reported a very low power devices of 4.8mW and also 1.5mW[ ],but in view of the very low cost design ,this is considered as a better option. The CC2500 was chosen because of its smaller size and lower power requirement. (Inspite of its limitations like : the CC2500 device suffered greatly from interference with other 2.4GHz systems, such as 802.11 wireless networks and Bluetooth and also the 2.4GHz band is greatly attenuated by the human body, limiting its suitability for body area networks).This transceiver is the major power consumer .In the present design it consumed nearly 61.9% of total power of 104mW.The MSP430 controller consumed nearly 21.3% of the total power when it is operated mostly in the mode LPM3 .The accelerometer consumption is around 3.1% of the total power i.e it consumed a power of 3.22mW.The other circuitry of the design consumed a power dissipation of about 4.5% of the total power.The entire power consumption is shown in Pi chart in the Figure 10.

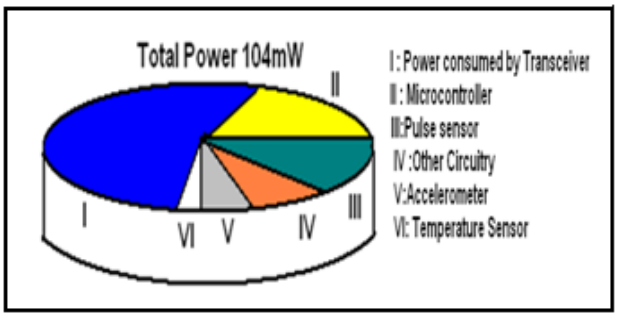


Figure 10. Power consumption in each subcomponent

After measuring the current consumption of different subcomponents of the present design ,it is observed that there is an advantage of nearly 10~11% power savings by this design when compared with normal design without low power modes of the microcontroller and with out application of TDMA based MAC protocols. The

Table.1 below shows the current/power consumed by each subcomponent of the present design.

Table 1.Details of Power consumption in the design

S.No	Sub Component	% of Power	Power consumed (mW)
I	Transceiver(WSN)	61.9	64.48
II	MSP430	21.3	22.25
III	Pulse Sensor	6.5	6.76
IV	Other circuitry	4.5	4.68
V	Accelerometer	3.1	3.22
VI	Temperature sensor	2.2	2.3

5.FUTURE SCOPE

In recent times ,because of inherent advantages of real time embedded systems ,researchers are more inclined towards the design of a real time wireless sensor designs. Currently the most important and widely used operating systems for wireless sensor networks are TinyOS ,Contiki and Mantis.The basic use of these operating systems is to provide a robust and reliable operation and maintain the system in the deepest low power mode so that the battery power life is extended. In view of these facts the authors are modifying the same design with MSP430 Microcontroller using the TinyOS .

6.ACKNOWLEDGEMENTS

One of the authors (Narasimha Murthy. Y) greatly acknowledges the financial support provided by the University Grants Commission, New Delhi,India, under Major research Project scheme.[F.NO.38-223/2009(SR)]

7.REFERENCES

[1] Johann Glaser and Daniel Weber and Sajjad A. Madani and Stefan Mahlknecht ,2008. Power aware simulation Frame work for Wireless Sensor Networks and Nodes, EURASIP Journal on embedded Systems ,2008.

[2] Atmel Corporation, ATmega 128L 8-bit AVR Low-Power MCU, Tech. Report, 2009.

[3] Bahareh Gholamzadeh and Hooman Nabovati,2008. Concepts of Designing Low Power Wireless Sensor Networks,World Academy of Science,Engineering and Technology ,21,pp 559-565 ,(2008).

[4] www.ti.com/msp430

[5] Johan Pouwelse ,Koen Langendoen and Henk Sips,2001.Dynamic Voltage Scaling on a Low-Power Microprocessor,ACM, ,pp 251-259(2001).

[6] Kazem Sohrby ,Daniel Minoli and Taieb Znati ,Wireless Sensor Networks : Technology, Protocols and Applications ,John Wiley @Sons, ,pp75-90, 2007

[7] V.Raghunathan, S.Ganeriwaland Mani.B.Srivastava, Emerging Techniques for long lived Wireless sensor Networks,IEEE Communication Magazine ,pp108-114, (2006).

[8] Mohammad Iliyas and Imad Mahgoub,Handbook of sensors networks :compact wireless and wired sensing systems ,CRC PressLLC,pp 125-131, 2005.

[9] Chipcon products from Texas Instruments,CC2500, Prelimanary Data Sheet (Rev.1.2) SWRS040a, pp1-84 .

[10] Javier Moreno Molina , Jan Haase and Christoph Grimm ,Energy Consumption Estimation and Profiling in Wireless Sensor Networks, private communication SNOPS Project 2010.

[11] R.Ramanathan and R.Rosales-Hain,2000.Topology Control of Multi hop Wireless Networks using Transmit power adjustment, INFOCOM , pp404-413, (2000)

[12]W.R.Heinzelman,A.Chandrakasan and H.Balakrishnan2000. Energy –Efficient Communication Protocols for Wireless Micro sensor networks ,Hawaii International Conference on System Sciences, 2000.

[13] V. Kumar, S. Sonavane and B. P. Patil,2009. Designing Ultra Low Power Wireless Sensor Network with TCP/IP link ,IEEE, 2nd International Conference on Adaptive Science & Technology, pp86-91, 2009.

[14] R. Lenhardt and D.I.Sessler,2006, Estimation of mean body temperature from mean skin and core temperature, J.of Anesthesiology,Vol.105,Dec.2006,pp1117-1121.

[15]http://intranet.daiict.ac.in/~ranjan/esp2006/lab/Tmp100.pdf

[16] Karandeep Malhi ,Subhas Chandra Mukhopadhyay,2012. A Zigbee based wearable Physiological Parameters Monitoring System, IEEE Sensors ,Vol.12,March 2012.

[17] Dogan Ibrahim, Kadri Buruncuk , Heart rate Measurement from the finger using a low cost Microcontroller, private communication, freedownloadb.com/pdf/heart-beat-monitoring-using-microcontroller.

[18] Andrew Bates, Martin Ling et al,2011. Accelerometer-based respiratory measurement during speech, International Conference on Body Sensor Networks.2011,pp95-100.

[19]https://www.sparkfun.com/datasheets Accelerometers/MMA7260Q-Rev1.pdf

[20] Hye-Sue Song and Paul M. Lehrer,2003.The Effects of Specific Respiratory Rates on Heart Rate and Heart Rate Variability, Applied Psychophysiology and Biofeedback, Vol. 28, March 2003 ,pp13-23

[21] ZigBee Aliance. ZigBee Specification,(2008).

http://www.zigbee.org/.

[22] http://www.iar.com/

[23] http://www.flashmagictool.com/

[24] A Cebrian,J.Rey,A.Torrnos and J.Millet ,2005.Adapting power consumption to performance requirements in a MSP430 microcontroller , Spanish Conference on Electron Devices ,IEEE Explorer, ,pp83-86(2005)

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 548 - 551, 2013, ISSN: 2319–8656

Data Dictionary Classification using Multilayer Artificial Neural Network with New Error Metrics

Jitendra Nath Shrivastava
Singhania University,
Jhunjhunu, Pacheri Bari,
Rajasthan, India

Maringanti Hima Bindu
Department of Computer Science and Applications,
North Orissa University, India

Abstract: The ANN technique is inspired by biological neurons. In past, Artificial Neural Networks has been used as a data classification technique. In this paper, artificial neural network is used as a data classifier. Here, new error metric are considered in the data classification. The results presented in the paper, clearly shows that ANN can acts as a very good classifier with new error metrics.

Keywords: Data Dictionary; ANN; LMS; geometric error metric, Harmonic error metric.

1. INTRODUCTION

In many applications, data classification is very necessary. In past Artificial neural network has been used as data classifier with accuracy more than 95%.[1,2] However, in some application the accuracy level of more than 99% is required. One such application is spam e-mail filtering where, data dictionary which consists of spam words need to be very accurate.

A neural network is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach [3]. Artificial neural networks are parallel computational models which are able to map any nonlinear functional relationship between an input and an output hyperspace to desired accuracy. They are constituted by individual processing units called neurons or nodes and differ among each other in the way these units are connected to process the information and, consequently in the kind of learning protocol adopted [4].

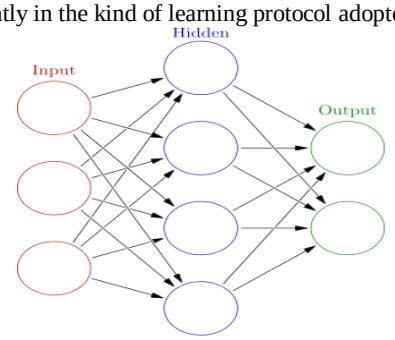


Figure 1 The Basic ANN Structure

In particular, the neurons of a feed-forward neural network are organized in three layers: the input units receive information from the outside world, usually in the form of a data file; the intermediate neurons, contained in one or more hidden layers, allow nonlinearity in the data processing, the output layer is used to provide an answer for a given set of input values [Figure 1]. In a fully connected artificial neural network, each neuron in a given layer is connected to each neuron in the following layer by an associated numerical weight (w_{ij}), the weight that passes between them. In addition, each neuron possesses a numerical bias term corresponding to an input of −1 whose associated weight has the meaning of a threshold value. Rumelhart et al. [5] popularized the use of back-propagation for learning internal representation in neural networks. Back-propagation (BP) algorithm is the most widely used search technique for training neural networks. Information in an ANN is stored in the connection weights which can be thought of as the memory of the system. The purpose of BP training is to change iteratively the weights between the neurons in a direction that minimizes the error E, defined as the squared difference between the desired and the actual outcomes of the output nodes, summed over training patterns (training set data) and the output neurons. The algorithm uses a sample-by-sample updating rule for adjusting connection weights in the network. In one algorithm iteration, a training sample is presented to the network. The signal is then fed in a forward manner through the network until the network output is obtained. The error between the actual and desired network outputs is calculated and used to adjust the connection weights.

Basically, the adjustment procedure, derived from a gradient descent method, is used to reduce the error magnitude. The procedure is firstly applied to the connection weights in the output layer, followed by the connection weights in the hidden layer next to output layer. This adjustment is continued backward through to network until connection weights in the first hidden layer are reached. The iteration is completed after all connection weights in the network have been adjusted. In this study, training of the multi-layer neural networks is implemented with back-propagation algorithm and network structure that has been trained with back-propagation algorithm has been used in the solutions of the multi-group classification models.

In Back-propagation algorithm, training of the neuron model is done by minimizing the error between target value and the observed value. In order to determine error between target and observed value, distance metric is used. It has been observed that Euclidean distance metric is the most commonly used for error measures in Neural Network applications. But it has been suggested that this distance metric is not appropriate for many problems [6]. In this work the aim is to find best error metric to use in Back propagation learning algorithm. The likelihood and log-likelihood functions are the basis for deriving estimators for parameters, for given set of data. In maximum likelihood method we estimate the value of 'y' for the given value of 'x' in presence of error. (See equation (1))

$$y_i = x_i + e \qquad (1)$$

Let x_i denote the data points in the distribution and let N denotes the number of data points. Then an estimator $\overline{\mu}$ of μ can be estimated by minimizing the error metric with respect to $\overline{\mu}$.

$$\varepsilon = \sum_{i=1}^{N} f(x, \overline{\mu}). \quad (2)$$

Where $f(x, \overline{\mu})$ is distance metric.

Jie, et. al., [7-8] has proposed some new distance metrics based on different means. These distance metrics can be used to improve the performance of the neuron model for learning the best-fit weights of the neuron models. Distance metrics associated with the distribution models that imply the arithmetic mean, harmonic mean and geometric mean in (See Table 1) are inferred using equation

$$\frac{d\varepsilon}{d\overline{\mu}} = \frac{d}{d\overline{\mu}} \sum_{i=1}^{N} f(x, \overline{\mu}) = 0 \quad (3)$$

In past, it is found that in the distribution associated with the harmonic and geometric estimations, the observations x_i which are far away from $\overline{\mu}$ will contribute less towards μ, in contrast to arithmetic mean and thus the estimated values will be less sensitive to the bad observations (i.e., observation with large variance), and therefore they are more robust in nature [9].

Table 1 Error Metric Types and Mean

	Error metric	Mean
Arithmetic	$\varepsilon = \sum_{i=1}^{N} (x_i - \mu)^2$	$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
Harmonic	$\varepsilon = \sum_{i=1}^{N} x_i \left(\frac{\mu}{x_i} - 1 \right)^2$	$\mu = \frac{N}{\sum_{i=1}^{N} x_i}$
Geometric	$\varepsilon = \sum_{i=1}^{N} \left[\log\left(\frac{\mu}{x_i} \right) \right]^2$	$\mu = \frac{1}{N} \left(\Pi_{i=1}^{N} x_i \right)^{1/N}$

Considering three layer structure of ANN, It has n_i, n_h and n_o neurons in input, hidden and output later respectively. Le the input and output vectors are $X = [x_1, x_2, ...x_{ni}]^T$ and $Y = [y_1, y_2, ...y_{ni}]^T$ respectively. Considering that the weight of the neuron that connects the i^{th} neuron of input layer to j^{th} neuron of output layer is w_{ij} and bias of the

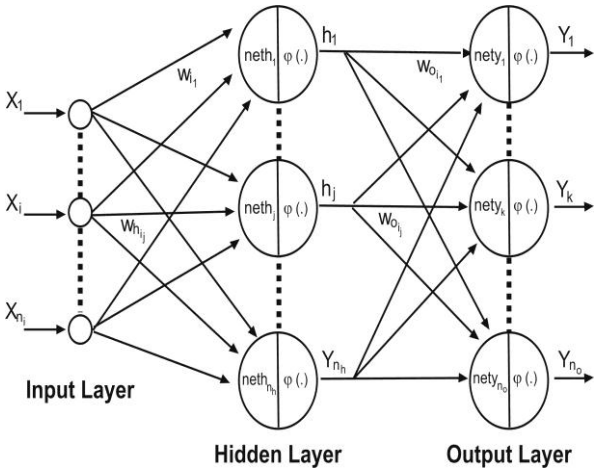


Figure 2 The Multilayer ANN Structure

j^{th} neuron of the hidden layer is bh_j, the net value of the j^{th} neuron is given by

$$neth_j = \sum_{i=1}^{n_i} wh_{ij}.x_i + bh_j \text{ where, } j = 1, 2...n_h. \quad (4)$$

The output of the j^{th} neuron h_j of the hidden layer after applying the activation is defined as

$$h_j = \varphi(neth_j) = \frac{1}{1 + e^{-neth_j}}. \quad (5)$$

Similarly, the net value $nety_k$ and the final output y_k of the k^{th} neuron of the output layer can be defined as

$$nety_k = \sum_{i=1}^{n_i} wo_{ij}.h_j + bh_k \text{ where, } k = 1, 2...n_o. \quad (6)$$

The output of the k^{th} neuron of the output layer

$$y_k = \varphi(nety_k) = \frac{1}{1 + e^{-nety_k}} \quad (7)$$

In this section the error back-propagation learning of MLP with different error metrics have been derived. Let E denote the cumulative error at the output layer. In BP algorithm aim is to minimize the error at the output layer. The weight update equations using gradient descent rule are given bellow:

$$wh_{ji}(n) = wh_{ji}(o) + \eta \frac{\partial E}{\partial y_k}.\frac{\partial y_k}{\partial wh_{ji}}$$

$$bh_j(n) = bh_j(o) + \eta \frac{\partial E}{\partial y_k}.\frac{\partial y_k}{\partial bh_j}$$

$$wo_{kj}(n) = wo_{kj}(o) + \eta \frac{\partial E}{\partial y_k}.\frac{\partial y_k}{\partial wo_{kj}}$$

$$bo_k(n) = bo_k(o) + \eta \frac{\partial E}{\partial y_k}.\frac{\partial y_k}{\partial bk_k} \quad (8)$$

Where, η is learning rate and

$$\frac{\partial y_k}{\partial wh_{ji}} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.w_{kj}\right].(1-h_j).h_j.x_i$$

$$\frac{\partial y_k}{\partial bh_j} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.w_{kj}\right].(1-h_j).h_j$$

$$\frac{\partial y_k}{\partial bo_k} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k\right]$$

$$\frac{\partial y_k}{\partial wo_{jk}} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.h_j\right] \quad (9)$$

For different error criterion only $\frac{\partial E}{\partial y}$ will change and is shown as follows: This is dependent on the distance metric used in computation of the total error E.

Case 1 Least Mean Square Error

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}(t_k - y_k)^2 \text{ , then}$$

$$\frac{\partial E}{\partial y_k} = -\eta(t_k - y_k) \quad (10)$$

Case 2 Geometric Error Metric

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}\log\left(\frac{y_k}{t_k}\right)^2 \text{ , then}$$

$$\frac{\partial E}{\partial y_k} = -\eta(\log(t_k) - \log(y_k).\frac{1}{y_k} \quad (11)$$

Case 3 Harmonic Error Metric

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}\left(\frac{y_k}{t_k} - 1\right)^2 \text{ , then}$$

$$\frac{\partial E}{\partial y_k} = -\eta.t_k.\left(\frac{y_k}{t_k} - 1\right) \quad (12)$$

Where, y_t denotes the desired value of neuron and t_i denotes the targeted value of the i^{th} pattern.

2. RESULTS

In our experiment 421 words are considered (appendix A). These words are classified into seven groups. The group division is done in the broad categories: adult, financial, commercial, beauty and diet, travelling, home based and gambling. The ANN network is trained with 1200 words. Then the ANN used as classifier with different error metric presented in Table1. The correct classification data in percentage is shown in Table 2. It can be observed for the table that, arithmetic error type produces 96% accurate results while geometric error type produces 98% accurate results. Among the three, harmonic error metric produces 99% correct results.

Table 2: Percentage Classification under Various Error Matrixes.

Type	Correct classification in percentage
Arithmetic	96%
Geometric	98%
Harmonic	99%

3. CONCLUSIONS

In this paper, ANN based data classifier is detailed with new error metric. Here, the wordlist we call it as data dictionary 421 words are considered as we use ANN based approach to classify them into seven categories. It has been fund that harmonic error metric data classification accuracy is 99%.

4. REFERENCES

[1] S. Haykin, Neural Networks: A Comprehensive Foundation, MacMillan College Publishing Company, New York, 1995.

[2] Kanellopoulos, I., Varfiss, A., Wilkinson, G.G. and Megier, J., 1991: Neural network classification of multi-data satelite imagery. Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS'91), Espoo, Finland, June, pp. 2215-2218.

[3] Rich Drewes "An artificial neural network spam classifier", Project home page: www.interstice.com/drewes/cs676/spam-nn

[4] Sexton, S.R., Dorsey, R.E., "Reliable classification using neural networks: a genetic algorithm and back-propagation comparison", ***Decision Support Systems***, 30: 11–22 (2000).

[5] Rumelhart, D.E., Hinton, G., Williams, R., "Learning representation by back-propagation errors", ***Nature***, 323(9): 533-536 (1986).

[6] W. J. J. Rey, "Introduction to Robust and Quasi-Robust Statistical Methods", Springer-Verlag, Berlin, 110–116, 1983.

[7] J. Yu, J. Amores, N. Sebe, Q. Tian, "Toward an improve Error metric", International Conference on Image Processing (ICIP), 2004.

[8] J. Yu, J. Amores, N. Sebe, Q. Tian, "Toward Robust Distance Metric Analysis for Similarity Estimation", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[9] [9] R. Battiti, "First and second-order methods for learning: between steepest descent and Newton's method", Neural Computation, Vol. 2, 141-166, 1992.

Group	Content	Example of Keywords in Each Group
C1	Adult	adult, aphrodisiac, big, cam, climax, company, cum, desire, erotic, fantasy, fuck, gay, girl, greate, guy, hard, hardcore, heaven, hot, huge, long, man, max, maxlength, nude, orgasm, penis, performance, pheromone, pill, porn, powerful, pussy, satisfy, sex, stamina, sweet, teen, viagra, webcam, x, xxx, xxx-porn, young, love, teen, anus
C2	Financial	Account, accountant, alert, analyst, attorney, bank, bankruptcy, benefit, bill, billing, broker, budget, building, cash, cheque, commission, consolidate, court, credit, creditor, currency, customer, debt, deposit, discover, economy, entrepreneur, estate, exchange, fee, finance, freedom, fund, help, high-risk, insurance, invest, investor, judgment, legal, legitimate, lender, loan, mastercard, mortgage, obligate, pay, payable, payable, paycheck, promote, purchase, rate, refinance, refund, rent, revenue, risk, service, statement, stock, support, tax, transaction, vat, visa, wealth, worth, service
C3	Commercial	college, commerce, computer, cost, deliver, discount, especial, expensive, express, fantastic, free, furnishing, furniture, game, get, gif, gift, great, guarantee, inexpensive, invite, item, just, keyboard, license, lifetime, magazine, maintenance, mall, market, material, materials, mobile, motherboard, mouse, offer, online, only, order, palm, pamphlet, percent, premium, price, produce, product, program, recommend, refill, release, resell, reseller, retail, sale, save, save, sell, ship, shipping, shop, shopping, special, subscribe, supply, surprise, trade, trademark, upgrade, voucher, whole, wholesale, within
C4	Beauty and Diet	after, age, amaze, anti-aging, appetite, beauty, become, before, believe, blood, body, botanic, breast, build, burn, Diet calorie, capsule, card, cell, change, chemical, cholesterol, confirm, course, diet, difference, dose, drug, effect, effective, eliminate, energy, enhance, exercise, eye, face, fast, fat, firm, fit, fitness, flexible, gary, grow, grown, growth, hair, health, healthcare, heart, height, herb, herbal, hormone, improve, inche, incredible, kidney, large, laser, life-changing, light, lose, loss, low, magic, medicine, metabolism, micro-cap, miracle, modem, move, muscle, nature, nutrient, old, over, overweight, permanent, plain, potential, pound, power, protect, reduce, remanufacture, repair, restore, retain, reverse, safe, satisfaction, secret, size, step, strength, strong, tablet, therapy, thin, toxin, treatment, under, virginia, vitamin, weight, woman, wonderful, wrinkle
C5	Traveling	book, deluxe, excite, guide, holiday, honest, hotel, luxury, meal, package, plan, problem, relax, relief, reserve, resort, summer, temple, ticket, tour, train, travel, traveler, trip, vacation,
C 6	Home-Based	address, astonishment, base, broadcast, bulk, business, comfort, connect, demo, domain, downline, download, Business earn, email, emailing, ethernet, facemail, fresh, home, homebased, homeworker, host, income, interest, international, internet, investigate, job, list, lucrative, mail, mailbox, mailer, mailing, make, marketing, message, million, money-making, opportunity, part-time, people, private, profit, reach, receive, recipient, require, re-register, return, server, software, subscriber, success, teach, unsubscribe, user, visit, website, work, work-at-home, worker, working
C7	Gambling	action, award, bet, bonus, casino, challenge, extra, gambling, gold, hunt, las, lucky, millionaire, player, poker, prize, reward, rich, vegas, win, lottery

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 552 - 556, 2013

Detection of Outliers and Reduction of their Undesirable Effects for Improving the Accuracy of K-means Clustering Algorithm

Bahman Askari
Department of Computer Science and Research Branch, Islamic Azad University, Khouzestan, Iran

Sattar Hashemi
Department of Computer Science & Engineering, Shiraz University, Shiraz, Iran

Mohammad Hossein Yektaei
Department of Computer Abadan Branch, Islamic Azad University, Abadan, Iran

Abstract: Clustering is an unsupervised categorization technique and also a highly used operation in data mining, in which, the data sets are divided into certain clusters according to similarity or dissimilarity criterions so that the assigned objects to each cluster would be more similar to each other comparing to the objects of other clusters. The k-means algorithm is one of the most well-known algorithms in clustering that is used in various models of data mining. The k-means categorizes a set of objects into certain number of clusters. One of the most important problems of this algorithm occurs when encountering to outliers. The outliers in the data set lead to getting away from the real cluster centers and consequently a reduction in the clustering algorithm accuracy. In this paper, we separate outliers from normal objects using a mechanism based on dissimilarity of objects. Then, the normal objects are clustered using k-means algorithm process and finally, the outliers are assigned to the closest cluster. The experimental results show the accuracy and efficiency of the proposed method.

Keywords: clustering, k-means, outliers, outlier detection

1. INTRODUCTION

Clustering is one of the highly used methods in data mining [12], wireless sensor networks [8,13], pattern recognition [14] and machine learning [7], which is used to detect the groups that are different enough from each other and contain similar objects [4,5]. The importance of clustering in various fields and also the type of data being used, clustering speed, accuracy and lots of other parameters, leads to introduce various methods and algorithms in data clustering. Clustering is an unsupervised technique in which the data sets which are usually vectors in multi-dimension space are divided into a certain number of clusters based on a similarity or dissimilarity criterions. For example, if the number of clusters is K, and there exist n number of m-dimension data, the clustering algorithm will assign each one of these data to a cluster. This assignment takes place according to this rule that the assigned data to a certain cluster are more similar to each other rather than the other clusters. The k-means algorithm is one of the most well-known clustering algorithms and is being used in various types of data mining. The k-means categorizes data set objects in certain numbers of clusters [10,3]. This method is one of the most attractive and highly used operations in clustering techniques, because it is simple and understandable and its time complexity is linear. In general, this algorithm consists of two phases. In the first phase, k numbers of objects are selected from data set in a random manner and are considered as the initial centers of clusters. In the second phase, the distance between objects and the clusters centers is determined and each object is placed in the nearest cluster. To determine the distance between objects, the Euclidean distance criterion is used, generally. When all of the objects have been placed in the corresponding clusters, the clusters centers are calculated using repetitive averaging of objects of each cluster. The second phase continues until satisfying the algorithm ending condition. The Pseudo-code of k-means algorithm is shown in figure 1.

K-means algorithm

Input: Data set D = { d_1, d_2, … d_n }, where d_i =data points, n= number of data points
K = number of cluster centers
Output:
Clusters : K clusters with their centers

Step 1:
Randomly select k data object from dataset D as initial cluster centers.
Step 2:
Repeat step 3 to step 4 till no new cluster centers are found
Step 3:
Calculate the distance between each data object di (1<=i<=n) and all K cluster centers Cj (1<=j<=K) and assign data object di to the nearest cluster.
Step 4:
For each cluster j (1<=j<=K), recalculate the cluster center.

Figure 1. The pseudo-code of k-means algorithm

The distance to the cluster center is calculated in the following method:

$$Distance(d_i, C_j) = \sqrt{\sum_{p=1}^{m} d_{ip} - C_{ip}} \quad (1)$$

Where, d_i denotes the i-th vector of data, C_j is the center of j-th cluster and m denotes the number of attributes and cluster centers.

The cluster centers are updated according to the following equation:

$$C_j = \frac{1}{n_j}\left[\sum_{\forall d_i \in Cluster_j} d_i\right] \quad (2)$$

Where n_j denotes the number of vectors in the j-th cluster and $cluster_j$ is a subset of all vectors which form the j-th cluster.

Figure 2 illustrates the clustering steps of a manual data set. This dataset is divided into two clusters using the k-means algorithm. At first, two clusters are formed by random selection of objects (figure 2-a), then the clusters centers are determined by averaging of objects of each cluster (figure 2-b) and the clustering process continues using the new cluster centers (figure 2-c). For this dataset, the clustering is finished after two iterations of the algorithm (figure 2-d).

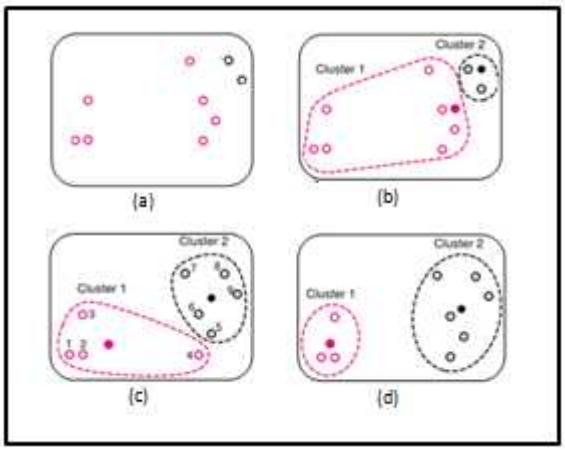


Figure 2. clustering of manual data using the k-means algorithm

The k-means algorithm has some pitfalls. For instance, it stops in local optimums and is sensitive to the initial values of clusters centers and outliers in dataset. In each dataset, an outlier is an object which its distance is not normal comparing to the other objects. In other words, an outlier is an object that has less similarity than the other objects. The existence of these objects in dataset has an undesirable effect on the efficiency and accuracy of the clustering algorithms such as the k-means.

2. RELATED WORK

Outlier detection has been a very interesting topic for research community [11,1,6,2,9]. Ramaswamy et al proposed a distance based outlier detection method in [11]. According to which, given parameters k and n, an object is an outlier if no more than n-1 other object in the dataset have higher value for D_k than object o, where $D_k(o)$ denotes the distance of k-th nearest neighbor of object o. This idea is further extended in [14], where each data point is ranked by the sum of distance from its k-th nearest neighbors. Breunig et al introduced the notion of the Local Outlier Factor (LOF) in [1] which captures the relative degree of outlierness of an object. It is local in sense that the degree of outlierness depends on how isolated an object is with respect to surrounding neighborhood. Above described methods are either distance based or nearest neighbors based that are not suitable for outlier detection in data streams due to their high time complexity. He et al in [6] presented new definition of outlier which they named as cluster-based local outlier, which provides importance to the local data behavior. They defined Cluster-Based Local Outlier Factor (CBLOF), a measure for identifying the physical significance of an outlier and an algorithm for discovering outliers is also proposed by them. After that Duan et al in [9] proposed a cluster based outlier detection algorithm which can detect both single point outliers and cluster-based outliers, and can assign each outlier a degree of being an outlier. Zhuo et al in [15] presented an outlier mining algorithm based on dissimilarity (OMABD), which detected outliers by comparing the dissimilarity degree with dissimilarity threshold. A dissimilarity based method is used for improving the efficiency of k-means algorithm in this study.

3. presented model

To improve the efficiency of the K-means algorithm, at first the dataset should be investigated for specifying and detecting the outliers. After that, the dataset should be divided into two subsets of normal object and outliers. Then, the clustering process should be performed separately for normal object and outliers. The normal object are clustered using the mentioned steps of the k-means algorithm shown in figure 1, and finally, the outliers are assigned to the nearest cluster according to the Euclidean distance criterion and calculated cluster centers from the previous steps.

3.1 Outlier detection in dataset

In this paper, the ODBD algorithm is presented for specifying and detecting the outliers which is based on the dissimilarity of objects. According to this algorithm, a value is calculated for all of the dataset objects which is called the dissimilarity degree. The objects that their degree is higher than the threshold value are considered as the outlier objects.

Assume that a data set DS is defined in the form of: DS= (D,A) in which $D=\{d_1,d_2,\dots d_n\}$ is the set of n objects and $A=\{a_1,a_2,\dots a_m\}$ is the set of attributes with the order of m. The dissimilarity degree of two objects d_i, $d_j \in D$ on the attribute of $f \in A$ is calculated as following:

$$ad_{ij}^{f} = \frac{\left(\left|d_{if} - d_f\right| - \left|d_{jf} - d_f\right|\right)^2}{d_{\max f} - d_{\min f}} \quad (3)$$

Where dif and d_{jf} are the values of attribute f in the objects i and j, respectively. d_f, d_{max} and d_{min} are the average, the maximum and the minimum value of attribute f on all objects of the dataset, respectively.

The dissimilarity degree of two objects can be obtained from the average of these objects dissimilarity on each of attributes, as following:

$$od(i,j) = \frac{\sum_{ak=1}^{m} ad_{ij}^{ak}}{m} \quad (4)$$

Where ad_{ij}^{ak} is the dissimilarity value of the objects i, j on the a_k-th attribute.

According to the relations (3) and (4), the dissimilarity matrix d_m, which is an order N square matrix, is created to calculate the dissimilarity of each object with respect to the other objects. By adding the row elements of this matrix, the objects synergic dissimilarity matrix is made which shows the dissimilarity degree of each object with respect to all other objects of the data set. For the sake of simplicity and simplification of comparisons, the synergic dissimilarity degree matrix is normalized and then, the average of elements of this matrix with an impact factor value in the range of [0,1] is considered as the threshold similarity value. The ODBD pseudo-code algorithm is shown in figure 3.

ODBD algorithm

input:
data set D = { $d_1, d_2, \dots d_n$ }, where d_i =data points, n= number of data points
impact factor value IFV ∈ [0,1]
output:
outlier and normal objects
step 1:
step 1-1:
for each data object d_i from D
  for each data object d_j from D
    calculate od(i,j) by using equation (3) and (4)
    dm(i,j)=od(i,j)
  end for
end for
step 1-2:
for each row r_i in dm
  calculate sum of elements and assign in sd_i
end for
calculate d_{max} = maximum value in sd
step 1-3:
for each data value sd_i in sd
  $dd_i = (d_{max} - sd_i) / d_{max}$
end for
td=mean(dd) *IFV
step 2:
for each data object d_i from D
  if dd_i < td
    assign d_i to outlier objects
  else
    assign d_i to normal objects
  end if
end for

Figure 3. The pseudo-code of ODBD algorithm

3.2 Improving the Clustering process with the ODBD-k-means algorithm

Selecting the initial values in the normal k-means algorithm is completely random. Thus, the existence of the outliers results in the cluster centers getting distance from real position and consequently decreasing the accuracy of this algorithm. In the presented algorithm it is attempted that the outliers do not affect the process of selecting the clusters center. For this purpose, after separating the data set objects using ODBD, the clustering is done during two phases. In the first phase, the normal process of k-means algorithm is used to cluster the normal object. In this phase, because of data set being pruned by the ODBD algorithm and the use of normal object, the centers and the objects are determined in a more accurate way. In the second phase, we use the centers obtained from the previous phase and with iteration, we calculate the distance between each of outliers and these centers. Then, each outlier is assigned to the nearest cluster. The Euclidean distance criterion is used for the calculation of this distance. The presented algorithm pseudo-code is illustrated in figure 4.

ODBD-K-Means algorithm

input:
data set D = { $d_1, d_2, \dots d_n$ }, where d_i =data points, n= number of data points
k = number of cluster centers
output:
k clusters with their centers
step 1:
find outliers and normal objects by using ODBD algorithm
step2:
step 2-1:
randomly select k data object from normal objects as initial cluster centers.
step 2-2:
repeat step 2-3 to step 2-4 till no new cluster centers are found
step 2-3:
calculate the distance between each data object d_i (1<=i<=size(normal object)) and all k cluster centers C_j (1<=j<=k) and assign data object d_i to the nearest cluster.
step 2-4:
for each cluster j (1<=j<=k), recalculate the cluster center.

step 3:
for each data object d_i from outliers
  step 3-1
  calculate the distance of d_i to all k final cluster centers C from step 2 by using euclidean distance
  step 3-2
  find the closest center c_j and assign d_i to the cluster with nearest center C_j
end for

Figure 4. The pseudo-code of ODBD-k-means algorithm

4. Results and discussion

To evaluate the algorithm presented in this study, the algorithm is implemented using MATLAB 2010 programming software and the results are compared with the k-means algorithm. The Iris, Bupa and Glass data sets from UCI are used in the experiments. The Iris data set is a categorization of iris flowers in which, there exists three different classes of iris and each class contains 50 objects. Each object has 4 attributes. In the Bupa data set, 345 objects exist each having 6 attributes. The attributes are gathered from blood tests concerning the diagnosis of liver hampering caused by irregular drink of alcohol. Each object of this data set is the record of a male person. In the glass data set, 214 objects exist and each object has 9 attributes and this set has 6 classes. The properties of these data sets are listed in table 1.

Table 1. The datasets and their properties

dataset	#objects	#features	#clusters
Iris	150	4	3
Bupa	345	6	2
Glass	214	9	6

Selecting the impact factor and consequently the similarity threshold value is very important to detect and determine the number of outliers. This factor value might be different for each data set. According to the ODBD algorithm, if the impact factor assumed to be zero, the number of outliers would be zero. This means that the ODBD-k-means algorithm changes to the normal k-means algorithm. For calculating the algorithm accuracy, we can use accuracy indicator which is obtained using to the following relation:

$$Accuracy = \frac{True\ Positive}{TruePositive + False\ Postive} \quad (5)$$

Because the number of normal object affects the selection of cluster centers and also the proposed algorithm like the k-means algorithm calculates the initial centers by random selection of objects, the results of each execution of this algorithm may not the same. For different values of impact factor in the range [0,1], the presented algorithm is executed for 100 times and according to the relevant number of outliers, the average value of results is considered as the algorithm accuracy. These results are presented in table 2, for Iris data set.

Table 2. Results of ODBD-k-means algorithm on the iris

Impact factor	#outlier	Accuracy
0.0	0	88.98
0.1	1	87.57
0.2	3	88.75
0.3	4	89.31
0.4	**5**	**90.36**
0.5	12	89.60
0.6	17	87.69
0.7	26	87.17
0.8	32	86.03
0.9	45	86.51
1.0	54	85.89

According to this table, it is obvious that the algorithm accuracy is dependent to the number of outliers. If we select zero as the value of the impact factor, the presented algorithm would be equivalent to the normal k-means algorithm. The real outliers of the Iris data set are the points that are separated with an impact factor of 0.4. By running the ODBD algorithm and selecting the most suitable impact factor for data set, the number of outliers in each data set is obtained. The results are shown in table 3.

Table 3. The number of outliers in data sets

dataset	Impact factor	#outlier objects
Iris	0.4	5
Bupa	0.8	12
Glass	0.6	8

Results of algorithm accuracy on the data sets are shown in table 4.

Table 4. The results of algorithms on datasets

Dataset	K-means	ODBD-K-means
Iris	88.98	**90.36**
Bupa	52.43	**53.72**
Glass	45.79	**47.09**

Figure 5 depicts the algorithm accuracy results on a diagram. According to this diagram, effect of the presented algorithm on the standard data sets can be observed.

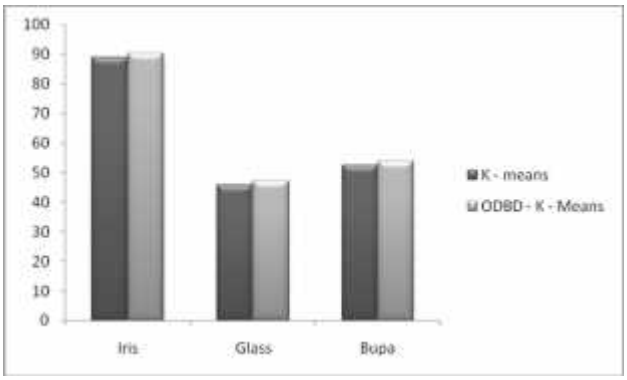


Figure 5. The results of algorithms on standard datasets

The k-means algorithm has a good accuracy on the Iris data set, because of a suitable distribution and structure of objects in data set. The existence of the outliers in the Bupa and Glass

data sets, has an undesirable effect on the selection of centers and objects and it consequently leads to reduction in the accuracy of the k-means algorithm. After temporarily removing this objects and using the presented algorithm, the retrieval accuracy on these three data sets have been improved.

5. Conclusion

In this paper, a new algorithm is presented based on the k-means algorithm and dissimilarity of the objects. In the presented model, the data sets were analyzed to specify the outliers. By detecting these objects, the data set pruned and the outliers and normal objects were separated from each other. Then, the normal objects were grouped using the k-means algorithm. Finally, using of the final cluster centers of previous stage and by calculating the distances between the outliers and the final cluster centers, these objects were assigned to the nearest cluster center, separately. The experimental results on the standard data sets showed that the presented algorithm probes the data sets with a better accuracy for finding better results. Despite the good results of the presented algorithm, different results were observed at different executions of the algorithm and that is due to the random selection of the initial points. In the future studies, it is possible to present better solutions for this challenge.

6. REFERENCES

[1] Breuing, M., Kriegel, H., Ng, R., and Sander J. 2000. LOF: identifying density-based local outlier factor. In Proceedings of ACM SIGMOD International Conference on management of Data., 29(2), pp. 93-104.

[2] Fabrizio, A., Stefano B., and Clara P. 2006. Distance-based detection and prediction of outliers. IEEE Transaction on Knowledge. And data Engineering., 18(2), pp. 145-160.

[3] Frogy, E. 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classification. Biometrics., 21(3), pp. 768-779.

[4] Guha, S., Rastogi, R., and Shim, K. 1998. An efficient clustring algorithm for large database. In Proceedings of the ACM SIGMOD Conference., 27(2), pp. 73-84.

[5] Guha, S., Rastogi R., and Shim K. 2000. A Robust clustring algorithm for categorical attributes. In Proceedings of the third IEEE International Conference on Data Mining., 25(5), pp. 345-366.

[6] He, Z., Xu X., and Deng, S. 2003. Discovering cluster-based local outlier. Pattern Recognization Letters., 24(9), pp. 1641-1650.

[7] Kao, Y., and Lee S.Y. 2009. Combining k-means and particle swarm optimization for dynemic data clustring problems. IEEE, International Conference on Intelligent Computing and Intelligent Systems., Shanghai, pp.757-761 .

[8] Kumar, M., and Verma, S. 2008. Data clustring in sensor network using art. 4th International Conference on Wireless Communication and Sensor Network, Allahabad, pp. 51-56.

[9] Lian D., Lida, X., Ying, L., and Jun, L. 2009. Cluster-based otlier detection. Annals of Operation Research, 68(11), pp. 151-168.

[10] Pelleg, D., and Moore, A. 2000. X-means: extending k-means with efficient estimation of the number of clusters. In Proceedings of ICML the Seventeenth International Conference on Machine Learning., San Francisco, pp. 727-734.

[11] Romaswany, S., Rastogi R., and Shim K. 2000. Efficient algorithm for mining outliers from large data set. In Proceedings of the ACM SIGMOD International Conference on Management of Data., Texas, ACM press, pp. 473-478.

[12] Tsai, C. F., Tsai, C. W., Wu, H. C. and Yang, T. 2006. A new data clustering approach for data mining in large databases. The 6th IEEE International Symposium on Parallel Architectures, Algorithms, and Networks., pp. 278-283.

[13] Wang, T., and Yang, Z. A location-aware-based data clustring algorithm in wireless sensor networks. 2008. IEEE, 11th International Conference on Communication Systems., pp. 1-5.

[14] Wong, A. K. C., and Li, G. C. L. 2008. simultaneous pattern and data clustring for pattern cluster analysis. IEEE Transaction on Knowledge and Data Engineering., 20(8), pp. 911-923.

[15] Zhou, M., and Chen, X. 2011. an outlier mining algorithm based on dissimilarity. International Conference on Environmental Science and Engineering., pp. 810-814.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 557 - 586, 2013, ISSN: 2319–8656

Analysis of Morphology Based Horticultural Features through Clustering Methods

K.Deb
Department of Computer Science & Engineering
Jadavpur University
Kolkata,India

A.Hazra
Department of Computer Science & Engineering
Jadavpur University
Kolkata,India

S.Kundu
Department of Computer Science & Engineering
Jadavpur University
Kolkata,India

P.Hazra
Faculty of Horticulture,
Bidhan Chandra Krishi Viswavidyalaya,
Kalyani,Nadia,India

Abstract: Cluster analysis is a prime Pattern Recognition method used to categorize sample patterns in a population by means of forming different clusters by assigning cluster memberships to the sample patterns depending on the feature similarity relationship among different patterns. Patterns displaying dissimilar feature values are assigned different cluster memberships whereas patterns carrying similar feature values are placed into same cluster. Searching the relationship among horticultural data has become a major research area in Pattern Recognition. In this paper we have used the morphological features for describing the characteristics of Tomato leaves and fruits belonging to different classes. Morphological feature values are extracted from different tomato leaf and fruiting habit samples to analyze through K-Means and Two-step clustering techniques to segment leaf and fruit samples into separate clusters according to their species owing to categorize them. Our experimentation also compares and discusses about the importance of the features which are obtained through K-Means and Two-step Clustering technique, may be useful for leaf and fruit species categorization.

Keywords: Cluster analysis, K-Means Clustering, Two-step clustering, Pattern Recognition, Horticulture, Morphological Feature.

1. INTRODUCTION

Economic growth of a nation depends highly on its agricultural and horticultural development. As different cultivars may need specifically different cultivation processes for better growth and quality development hence it is very much needed to identify the horticultural cultivars belonging to different classes independently, so that appropriate cultivation means can be applied for specific cultivar. Thus horticultural species categorization is a very important task for cultivation. This task of categorization is easy to perform for the crops with small number of species variations, but the task becomes a tough one if huge numbers of crop species are to be dealt with. Horticultural vegetable like tomato have large variety of species found all around the globe. The high variation of morphological feature values of tomato leaf and fruit among different tomato species is a prominent indicator of species diversity upon which cluster analysis would be applied to form different tomato-species clusters and thus categorizing the tomato species depending on their cluster memberships. This automatic method of species categorization through clustering exhibits high level of accuracy and requires trifle time compare to manual process. Cluster analysis also produces some distinguishing results through which the feature importance of leaves and fruits could be predicted which may be of very useful while classifying the particular plant species.

2. BASIC CLUSTER ANALYSIS

Cluster analysis is an important analytical procedure used for the purpose of analysing data. Cluster analysis is widely used in different research areas like machine learning, pattern recognition, market research, digital image processing, Biology etc. Basically, clustering divides the sample entities into different groups called clusters depending on similarity present between entities. Entities placed into same cluster, bear great deal of similar features where as entities belonging to different clusters don't have that much of feature similarity like same cluster entities. All the member entities of a cluster can be represented by the cluster centre of that cluster. Now with the compact cluster formation, obtaining information from the original entity set can be sufficiently reduced to collecting information about a small number of clusters. Information obtained from the clusters can be very effective for purposes like entity classification, identification etc.

Clustering algorithms can be classified into categories such as: Partitional Clustering and Hierarchical Agglomerative Clustering. We discuss these two clustering techniques briefly in the following-

Partitional Clustering starts with some initial clusters. For each of the initial clusters, a cluster center is calculated by fulfilling the optimality condition. Sample objects are placed in different clusters depending on the smallest distance criterion i.e. a sample object is placed in that cluster whose cluster center is minimum distance away from the sample object. Sample input data are partitioned into the initial clusters. In the next step, cluster centers are recalculated and objects are again placed in different clusters depending on the new calculated cluster centers. This process of cluster center

recalculation and placing the objects in clusters continues unless the placement of the objects in the clusters remains unaltered between two successive rotations.

Hierarchical Agglomerative Clustering algorithm starts with some single clusters depending on the size of the input data set. Number of initial single clusters is equal to the size of the input data set and each of the input patterns belongs to different cluster. Now as the algorithm moves, at each of the successive steps, merging of the cluster pairs having highest level of attribute similarities is performed.

3. APPLIED CLUSTERING METHODS

3.1 K-means Cluster Analysis

K-Means cluster analysis falls into the category of partional clustering algorithm. K-Means cluster analysis is used for analysing the feature data set.

Let's assume that P number of sample patterns is to be clustered. A pattern set $D=\{d_1,d_2,..,d_P\}$represents the sample patterns .The characteristics of each sample pattern is represented by Z number of features, which constitute the feature set $F=\{f_1,f_2,\ldots\ldots,f_Z\}$. Now for each of the Z features, P different feature values are obtained from each of the P sample patterns. The feature values associated to a feature f_X forms the individual feature value set $IFV_X=\{ifv_{1X},ifv_{2X},...,ifv_{NX},...,ifv_{PX}\}$,of size P, where ' ifv_{NX}', an element of the set IFV_X, denotes the feature value of N th sample pattern with respect to feature f_X. K-Means cluster analysis is done on each IFV, to place P patterns in k(user given value) different clusters depending on the values of the elements of the IFV. K-Means clustering initially selects k patterns out of P patterns as initial clusters. Each cluster is represented by a cluster center. The value of each initial cluster center will be one of the elements of the IFV chosen randomly with uniqueness condition that same IFV element can't be placed into more than one initial clusters. Also the cluster membership of a particular IFV element remains same till the end of the clustering process.

Let's consider that K-Means clustering is applied to IFV_X. This will lead to the formation of k clusters each having a cluster center. Let's consider that the cluster center of 'i' th cluster is denoted by CC_i. Let's denote the value of cluster center CC_i by VCC_i. Now 'M' th data pattern d_M will be placed into the 'i' th cluster by satisfying the condition $Dis(d_M,CC_i) < Dis(d_M,CC_j)$,for all $j\neq i$, where $Dis(d_M,CC_i)$ is the distance between the data pattern d_M and the 'i' th cluster center CC_i and $Dis(d_M,CC_j)$ is the distance between d_M and another 'j' th cluster center CC_j. Now, $Dis(d_M,CC_i)$ can be calculated as per the following equation-

$$Dis(d_M,CC_i) = | ifv_{MX} - VCC_i | \quad (1)$$

The values of k cluster centers will be recalculated again and again unless no new member is placed in clusters. The updated value of a cluster center is the calculated average of the values member elements of the cluster. So VCC_i is updated as -

$$VCC_i = (vme_{1i} + vme_{2i} + ...+ vme_{li} +\ldots + vme_{si}).(1/s) \quad (2)$$

, where 'vme_{1i}' denotes the value of 'l' th member element of 'i' th cluster having 's' number of member elements of the 'i' th cluster. Basically value of each member element is an element of set IFV_X.

In the following, we summarize different steps of K-Means algorithm done on the set IFV_X -

1) Randomly choose k number of initial cluster centers out of P elements of the IFV_X set.
2) Place each pattern from pattern set D in the cluster whose cluster center is closest to it by calculating the pattern-cluster distance as per equation (1).
3) Recompute cluster centers as per equation (2), depending on the recent placement of the elements into the cluster and reassign the elements to its closest cluster based on the newly computed centers.
4) Repeat step 2 and 3 until there is no alteration in the cluster memberships.

3.2 Two step Cluster Analysis

Two-step cluster analysis belongs to the class of Hierarchical Agglomerative Clustering. Consider the pattern set D of size P and feature set F of size Z as mentioned in section 3.1. Values of Z features are extracted from each pattern. Hence values related to Z features, extracted from I th pattern d_I, forms the values of features set $VF_I=\{vf_{I1},vf_{I2},\ldots..,vf_{IJ},\ldots.vf_{IZ}\}$,where vf_{IJ} is the value of J th feature extracted from I th pattern. Each pattern is represented by its VF set. Now the values of all features extracted from all P patterns build a set of all values of features $AVF=\{VF_1,VF_2,\ldots VF_I\ldots.,VF_P\}$. Two-step cluster analysis is performed on the set AVF .Two-step clustering initially forms some sub-clusters and places P patterns into them. Each pattern is described by its VF values, so each sub-cluster contains the VF value of the member pattern. Let's consider that sc_A and sc_B are two sub-clusters containing pattern d_I and d_K respectively. Now the distance between sc_A and sc_B is calculated by calculating the Euclidean Distance between VF_I and VF_K, denoted by $ED(VF_I,VF_K)$,in the following equation-

$$ED(VF_I,VF_K)=((vf_{I1} - vf_{K1})^2+(vf_{I2} - vf_{K2})^2+..+(vf_{IZ} - vf_{KZ})^2)^{1/2} \quad (3)$$

The Two-step algorithm operates on AVF set in the following manner-

1) Place all the P sample patterns into different sub clusters depending on the values of the features set (VF) of each pattern. So each sub-cluster contain the VF set of a pattern.
2) Calculate the distance between sub clusters using equation (3). Merge two nearest sub-clusters(clusters with minimum distance between each other) into one cluster to form new clusters.

Repeat the process of nearest cluster merging between the new clusters until desired number of clusters are formed.

4. EXPERIMENTAL RESULTS

In this paper we have used the best selected morphological features [1] to perform K-Means and Two-step clustering with the help of IBM SPSS statistics 20 data mining tool and there by comparing the cluster building abilities of these features. This comparison will give better visibility about the impact of the features in machine vision solutions. The morphological features used in our experiment are listed below -

Leaf Features

1) Major Axis
2) Minor Axis
3) Aspect Ratio
4) Eccentricity
5) Area
6) Rectangularity

7) Diameter
8) Compactness
9) Perimeter Ratio of Major Axis-Minor Axis
10) Perimeter Ratio of Diameter
11) Concavity
12) R-Factor

Fruit Features

1) Branch Length
2) Branch Width
3) Length Width Ratio
4) Area
5) Perimeter
6) Equivalent Diameter
7) Rectangularity
8) Diameter
9) Perimeter Ratio of Branch Length-Branch Width
10) Perimeter Ratio of Diameter
11) Convexity
12) Solidity
13) On Pixels
14) Narrow-Factor

4.1 K-Means Clustering Results

K-Means clustering is performed individually on feature values, related to each individual leaf and fruit features, extracted from all sample leaf and fruit patterns.

K-Means process starts by randomly selecting 15 and 14 initial cluster centers of individual leaf features (Table 1) and individual fruit features (Table 2) respectively.

Table 1- Randomly chosen Initial cluster centers for individual leaf features

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Major Axis	479	390	419	339	370	435	587	503

Initial Cluster Centers

	Cluster						
	9	10	11	12	13	14	15
Major Axis	599	533	519	575	621	654	557

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Minor Axis	167	321	216	197	254	182	279	291

Initial Cluster Centers

	Cluster							
	9	10	11	12	13	14	15	
Minor Axis	150	303	366	404	266	230	245	

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Aspect Ratio	2.523810	2.795276	3.086207	3.264000	3.175000	2.081340	3.462264

Initial Cluster Centers

	Cluster							
	8	9	10	11	12	13	14	
Aspect Ratio	1.912371	2.939850	4.360000	1.602459	1.423792	2.642336	2.410072	

Initial Cluster Centers

	Cluster
	15
Aspect Ratio	2.260116

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Eccentricity	.933819	.837966	.943856	.803081	.924178	.874209	.957381	.852387

Initial Cluster Centers

	Cluster							
	9	10	11	12	13	14	15	
Eccentricity	.885553	.824524	.781392	.711832	.893455	.914260	.973342	

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Area	279909.125	321362.000	315217.375	337626.625	343288.250	330031.625

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Area	298127.250	301354.125	271183.500	276812.625	290063.875	262710.500

Initial Cluster Centers

	Cluster		
	13	14	15
Area	285034.125	304995.625	295351.000

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Rectangularity	.525200	.252211	.287069	.202822	.320442	.369372	.551564	.426617

Initial Cluster Centers

	Cluster						
	9	10	11	12	13	14	15
Rectangularity	.621876	.586994	.792047	.883636	.504833	.463316	.403556

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Diameter	414	316	398	441	356	492	470	614

Initial Cluster Centers

	Cluster						
	9	10	11	12	13	14	15
Diameter	597	650	584	564	545	515	530

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Compactness	1.530433	2.035084	1.588200	1.205719	1.857097	1.249416	1.063003

Initial Cluster Centers

	Cluster						
	8	9	10	11	12	13	14
Compactness	1.350639	.982277	1.284001	1.439500	1.319113	1.104743	1.375947

Initial Cluster Centers

	Cluster
	15
Compactness	1.157990

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Perimeter Ratio of Major Axis-Minor Axis	6.694405	7.195364	7.080442	5.983089	7.969643	7.500000

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Perimeter Ratio of Major Axis-Minor Axis	5.771539	5.629520	5.404010	6.914864	5.087829	4.831944

Initial Cluster Centers

	Cluster		
	13	14	15
Perimeter Ratio of Major Axis-Minor Axis	6.293952	6.139844	6.528949

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Perimeter Ratio of Diameter	8.718986	13.753165	11.278894	13.200000	12.536517	8.997152

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Perimeter Ratio of Diameter	7.576137	7.902203	8.465625	9.509748	8.228941	10.629945

Initial Cluster Centers

	Cluster		
	13	14	15
Perimeter Ratio of Diameter	10.257154	9.872329	6.920409

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Concavity	91802.875	50350.000	56494.625	34085.375	28423.750	41680.375	73584.750

Initial Cluster Centers

	Cluster					
	8	9	10	11	12	13
Concavity	70357.875	100528.500	94899.375	81648.125	108918.500	86677.875

Initial Cluster Centers

	Cluster	
	14	15
Concavity	66716.375	76361.000

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
R-factor	897.855072	1176.303797	933.949749	1143.729231	1044.134831	745.879518

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
R-factor	791.156550	571.898383	622.190955	769.043478	658.595745	842.285714

Initial Cluster Centers

	Cluster		
	13	14	15
R-factor	682.181818	726.193548	703.500000

Table 2- Randomly chosen Initial cluster centers for individual fruit features

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Branch Length	508.861407	631.348827	484.844350	654.464819	574.908316	691.991471

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Branch Length	447.317697	598.925373	586.916844	537.381663	553.893390	615.437100

Initial Cluster Centers

	Cluster	
	13	14
Branch Length	520.869936	673.978678

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Branch Width	411.292111	358.754797	475.837953	321.228145	382.771855	295.710021

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Branch Width	447.317697	496.852878	460.827292	256.682303	424.801706	340.742004

Initial Cluster Centers

	Cluster	
	13	14
Branch Width	235.667377	306.217484

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Length width Ratio	2.491228	1.784404	1.017065	2.039409	1.491857	2.192308	1.206790

Initial Cluster Centers

	Cluster						
	8	9	10	11	12	13	14
Length width Ratio	1.414894	2.821656	2.319797	1.306569	1.650655	1.577061	1.926471

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Area	175316	207688	186587	240592	292808	269369	200254	227813

Initial Cluster Centers

	Cluster					
	9	10	11	12	13	14
Area	248800	134572	256387	193219	160739	213099

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Perimeter	4396	4808	4946	5793	5720	6427	5205	4603

Initial Cluster Centers

	Cluster					
	9	10	11	12	13	14
Perimeter	5056	5475	3815	5319	4530	5569

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
Equivalent Diameter	472.461197	514.234433	487.411966	553.472282	610.585559

Initial Cluster Centers

	Cluster				
	6	7	8	9	10
Equivalent Diameter	585.637892	504.946683	538.572670	562.834324	413.934674

Initial Cluster Centers

	Cluster			
	11	12	13	14
Equivalent Diameter	571.350950	495.998216	452.393334	520.889542

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Rectangularity	1.51555	.98869	1.35499	.56056	.74742	.89254	1.25091	1.46348

Initial Cluster Centers

	Cluster					
	9	10	11	12	13	14
Rectangularity	1.06472	1.12193	1.58324	1.20264	2.47732	1.41198

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Diameter	421.799574	483.343284	522.371002	570.405117	639.454158	675.479744

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Diameter	597.424307	619.940299	580.912580	661.970149	505.859275	697.995736

Initial Cluster Centers

	Cluster	
	13	14
Diameter	609.432836	447.317697

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Perimeter Ratio of Branch Length-Branch Width	4.430537	4.718868	5.148424	4.235480	5.946467	5.038155

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Perimeter Ratio of Branch Length-Branch Width	4.880447	3.996112	5.745399	5.286098	6.160609	3.301852

Initial Cluster Centers

	Cluster	
	13	14
Perimeter Ratio of Branch Length-Branch Width	5.415092	5.539335

Initial Cluster Centers

	Cluster					
	1	2	3	4	5	6
Perimeter Ratio of Diameter	10.422011	9.085497	8.155919	8.447897	9.686931	7.963229

Initial Cluster Centers

	Cluster					
	7	8	9	10	11	12
Perimeter Ratio of Diameter	9.372221	5.772676	8.703547	7.397360	8.894112	6.618560

Initial Cluster Centers

	Cluster	
	13	14
Perimeter Ratio of Diameter	7.707393	11.890878

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Convexity	84.55687	79.52760	77.31115	72.58582	64.13240	57.83600	70.96449

Initial Cluster Centers

	Cluster						
	8	9	10	11	12	13	14
Convexity	82.05563	80.74126	68.10407	66.74663	69.29754	87.08688	74.88155

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Solidity	.471645	.558735	.501968	.647254	.787728	.724672	.538734	.612875

Initial Cluster Centers

	Cluster					
	9	10	11	12	13	14
Solidity	.669336	.400704	.689746	.519809	.434213	.573290

Initial Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
On-Pixels	175227	207604	186498	240498	292731	269295	200173	227711

Initial Cluster Centers

	Cluster					
	9	10	11	12	13	14
On-Pixels	248725	134514	256313	193144	160675	212979

Initial Cluster Centers

	Cluster						
	1	2	3	4	5	6	7
Narrow-Factor	.891534	.819477	1.250774	.927954	.953964	.986842	1.000000

Initial Cluster Centers

	Cluster						
	8	9	10	11	12	13	14
Narrow-Factor	1.146341	1.015284	1.065491	1.038560	.974359	1.094017	1.111959

Table 3 and Table 4 represent the iteration history of the K-Means clustering on individual leaf and fruit features respectively. Iteration history shows the number of times the clustering iterates before completion. Clustering algorithm completes when there is a small change or no change in the cluster centers, there by achieving the convergence.

Table 3- Iteration History of clustering on 12 individual leaf features

Table 3.1- Iteration History of clustering on Major Axis

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.000	2.150	.000	.000	.000	.530	2.502
2	.000	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.000	3.002	.000	4.203	4.003	.000	4.000E-007
2	.000	1.126	.000	.000	.000	.000	1.334
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 12.009.

Table 3.2- Iteration History of clustering on Minor Axis

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.375	.000	.790	.461	1.501	.375	1.001	.000
2	2.377	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	3.753	.000	.000	.000	3.002	1.876	3.002
2	1.751	.000	.000	.000	.000	1.876	1.501
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 9.006.

Table 3.3- Iteration History of clustering on Aspect Ratio

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.021	.000	.015	.000	.016	.028	.013	.041
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.000	.000	.053	.000	.023	.005	.001
2	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .089.

Table 3.4- Iteration History of clustering on Eccentricity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.002	.001	.002	.001	.002	.002	.002	.004
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.002	.000	.000	.000	.003	2.325E-005	.000
2	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .008.

Table 3.5- Iteration History of clustering on Area

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	1280.688	804.875	561.531	.000	.000	1761.583	92.188	605.896
2	.000	.000	.000	.000	.000	.000	473.604	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	109.938	.000	477.083	.000	.000	49.875	788.925
2	.000	.000	.000	.000	.000	.000	513.044
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 2776.250.

Table 3.6- Iteration History of clustering on Rectangularity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.007	.002	.004	.001	.003	.001	.000	.006
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.004	.000	.000	.000	.006	.000	.007
2	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .020.

Table 3.7- Iteration History of clustering on Diameter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	4.500	.000	4.503	.000	1.629	6.004	3.002
2	.000	.000	.000	.000	.000	1.329	1.501	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.000	.000	4.878	1.001	.429	.500	1.876
2	.000	.000	1.126	4.003	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 13.510.

Table 3.8- Iteration History of clustering on Compactness

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.009	.000	.001	.000	8.000E-006	.005	.000
2	.000	.000	.000	.000	.000	.000	.008	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.017	8.050E-005	.000	.004	.001	.000	.002
2	.000	.000	.000	.000	.007	.000	.000
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any

center is .000. The current iteration is 3. The minimum distance between initial centers is .025.

Table 3.9- Iteration History of clustering on Perimeter Ratio of Major Axis-Minor Axis

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.034	.000	.000	.018	.020	.025	.039	.007
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.044	.000	.023	.000	.012	.027	.058
2	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .115.

Table 3.10- Iteration History of clustering on Perimeter Ratio of Diameter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.014	.000	.004	.000	.000	.080	.072	.013
2	.040	.000	.000	.000	.000	.026	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	.034	.008	.009	.000	.048	.028	.086
2	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is .237.

Table 3.11- Iteration History of clustering on Concavity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	1280.688	804.875	561.531	.000	.000	1761.583	92.188	605.896
2	.000	.000	.000	.000	.000	.000	473.604	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	109.938	.000	477.083	.000	.000	49.875	788.925
2	.000	.000	.000	.000	.000	.000	513.044
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 2776.250.

Table 3.12- Iteration History of clustering on R-Factor

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.000	.000	.000	.000	1.634	3.755	.000
2	.000	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers						
	9	10	11	12	13	14	15
1	3.273	1.367	7.984	8.423	.602	3.522	4.450
2	5.889	.000	2.004	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 19.686.

Table 4- Iteration History of clustering on 14 individual fruit features

Table 4.1- Iteration History of clustering on Branch-Length

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	1.501	.000	5.004	3.753	4.503	.000	1.501
2	.000	.000	.000	1.376	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.563	.000	2.502	3.002	3.002	3.002
2	.000	.000	.000	.000	.000	1.501
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 12.009.

Table 4.2- Iteration History of clustering on Branch-Width

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	5.254	4.003	3.333E-007	4.503	3.377	.000	3.603	6.004
2	.000	.000	3.753	.000	.000	.000	3.603	1.876
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	1.501	.000	.751	3.377	.000	1.501
2	1.501	.000	1.951	.000	.000	.000
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 10.507.

Table 4.3- Iteration History of clustering on Length-Width Ratio

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.021	.013	.004	.010	.000	.025	.023
2	.000	.000	.000	.000	.000	.000	.007	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.000	.000	.019	.004	.016	.021
2	.000	.000	.009	.000	.000	.000
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is .074.

Table 4.4- Iteration History of clustering on Area

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	649.813	420.825	1662.792	1128.833	.000	2781.417	.000	2060.188
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	529.025	.000	1237.800	.000	331.438	2533.313
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 5410.500.

Table 4.5- Iteration History of clustering on Perimeter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	31.000	9.000	.000	.000	.000	2.889	37.500
2	.000	.000	.000	.000	.000	.000	8.486	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	24.333	4.667	20.500	10.444	15.000	7.500
2	14.917	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 73.000.

Table 4.6- Iteration History of clustering on Equivalent-Diameter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.877	.516	2.185	1.309	.000	3.038	.000	2.422
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.804	.000	1.386	.000	.466	3.078
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 6.655.

Table 4.7- Iteration History of clustering on Rectangularity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.003	.016	.014	.013	.022	.002	.017	.000
2	.000	.006	.000	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.004	.012	.006	.004	.000	.001
2	.012	.000	.000	.000	.000	.000
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is .048.

Table 4.8- Iteration History of clustering on Diameter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.000	2.252	2.502	5.000E-007	1.876	3.002	.000
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.375	.188	.000	3.002	1.501	.000
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 10.507.

Table 4.9- Iteration History of clustering on Perimeter Ratio of Branch Length-Branch Width

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.039	.010	.000	.009	.000	.023	.000
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.007	.024	.024	.000	.013	.000
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .110.

Table 4.10- Iteration History of clustering on Perimeter Ratio of Diameter

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.014	.045	.049	.000	.049	.037	.000
2	.000	.000	.000	.000	.000	.016	.000	.000
3	.000	.000	.000	.000	.000	.015	.000	.000
4	.000	.000	.000	.000	.000	.012	.000	.000
5	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.016	.014	.015	.000	.009	.000
2	.000	.000	.000	.000	.015	.000
3	.000	.000	.000	.000	.016	.000
4	.000	.000	.000	.000	.021	.000
5	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 5. The minimum distance between initial centers is .191.

Table 4.11- Iteration History of clustering on Convexity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.529	.029	.505	.370	.426	.000	.096	.270
2	.000	.000	.000	.000	.000	.000	.199	.000
3	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.055	.268	.090	.302	.000	.136
2	.000	.000	.000	.148	.000	.000
3	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 1.193.

Table 4.12- Iteration History of clustering on Solidity

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.002	.001	.004	.003	.000	.007	.000	.006
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.001	.000	.003	.000	.000	.007
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .015.

Table 4.13- Iteration History of clustering on On Pixels

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	649.500	418.800	1671.667	1112.667	.000	2790.667	.000	2075.250
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	530.200	.000	1240.000	.000	316.000	2549.500
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 5375.000.

Table 4.14- Iteration History of clustering on Narrow-Factor

a
Iteration History

Iteration	Change in Cluster Centers							
	1	2	3	4	5	6	7	8
1	.000	.005	.000	.001	.000	.000	.001	.005
2	.000	.000	.000	.000	.000	.000	.000	.000

a
Iteration History

Iteration	Change in Cluster Centers					
	9	10	11	12	13	14
1	.001	.000	.005	.000	.006	.005
2	.000	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is .012.

Information regarding the membership of 15 leaf and 14 fruit clusters build through K-Means Clustering is shown by Table-5 and Table 6. The case number field signifies the sample leaf/fruit pattern number. The field called cluster is the cluster number in which a pattern is placed and distance field gives the distance between the pattern and the cluster center, in which the pattern is placed.

Table 5-Cluster membership of leaf clusters build from 12 individual leaf features

Table 5.1-Cluster membership for Major Axis

Cluster Membership			Cluster Membership		
Case Number	Cluster	Distance	Case Number	Case Number	Case Number
1	1	.000	16	10	.375
2	2	.000	17	15	7.339
3	3	2.150	18	13	1.001
4	4	.000	19	15	4.670
5	5	.000	20	12	1.201
6	6	.000	21	10	7.881
7	7	1.148	22	8	1.001
8	15	1.334	23	10	4.128
9	9	.000	24	15	1.668
10	10	1.126	25	15	.167
11	7	.530	26	11	.000
12	12	4.203	27	10	4.878
13	15	1.334	28	8	2.502
14	7	2.031	29	7	.530
15	15	1.668	30	12	4.803

Cluster Membership		
Case Number	Cluster	Distance
31	11	.000
32	13	4.003
33	7	2.472
34	12	4.203
35	7	.530
36	10	5.629
37	10	6.380
38	15	3.169
39	14	.000
40	3	2.150
41	12	4.803
42	10	8.631
43	8	3.502
44	15	1.334
45	13	5.004

Table 5.2-Cluster membership for Minor Axis

Cluster Membership		
Case Number	Cluster	Distance
1	3	2.637
2	3	1.363
3	3	.363
4	4	4.900
5	4	7.100
6	3	6.363
7	7	1.001
8	8	.000
9	7	.500
10	10	.000
11	11	.000
12	12	.000
13	13	3.002
14	15	3.002
15	15	4.503

Cluster Membership		
Case Number	Cluster	Distance
16	1	5.504
17	9	3.502
18	14	2.842E-014
19	6	.375
20	6	1.876
21	9	2.001
22	1	2.001
23	4	6.465
24	3	5.294
25	14	2.842E-014
26	3	.790
27	2	.000
28	3	6.715
29	4	2.542
30	1	3.502

Cluster Membership		
Case Number	Cluster	Distance
31	5	3.333E-007
32	5	1.501
33	5	1.501
34	14	2.842E-014
35	13	3.002
36	15	1.501
37	4	8.546
38	6	3.377
39	9	5.504
40	4	.461
41	3	.790
42	4	1.962
43	7	.500
44	3	5.294
45	6	5.629

Table 5.3-Cluster membership for Aspect Ratio

Cluster Membership		
Case Number	Cluster	Distance
1	15	.062
2	8	.049
3	8	.077
4	11	.022
5	8	.076
6	6	.028
7	6	.001
8	8	.041
9	6	.048
10	8	.104
11	11	.053
12	12	.000
13	6	.013
14	1	.034
15	15	.029

Cluster Membership		
Case Number	Cluster	Distance
16	3	.015
17	7	.013
18	13	.028
19	3	.028
20	5	.016
21	7	.008
22	3	.044
23	2	.000
24	1	.035
25	14	.017
26	14	.012
27	11	.031
28	14	.005
29	9	.000
30	7	.022

Cluster Membership		
Case Number	Cluster	Distance
31	6	.050
32	1	.023
33	15	.066
34	1	.001
35	15	.001
36	15	.033
37	13	.023
38	5	.016
39	10	.000
40	6	.044
41	13	.047
42	13	.042
43	8	.039
44	1	.021
45	4	.000

Table 5.4-Cluster membership for Eccentricity

Cluster Membership		
Case Number	Cluster	Distance
1	13	.006
2	2	.001
3	8	.002
4	4	.001
5	8	.002
6	6	.000
7	9	.004
8	8	.004
9	9	.002
10	10	.000
11	11	.000
12	12	.000
13	6	.002
14	14	2.325E-005
15	13	.003

Cluster Membership		
Case Number	Cluster	Distance
16	3	.000
17	7	.002
18	5	.002
19	3	.001
20	3	.003
21	7	.001
22	3	.002
23	1	.002
24	14	.005
25	14	.003
26	14	.005
27	4	.001
28	14	.004
29	3	.005
30	7	.001

Cluster Membership		
Case Number	Cluster	Distance
31	6	.002
32	14	.001
33	13	.006
34	14	.003
35	13	.000
36	13	.003
37	5	.000
38	3	.002
39	15	.000
40	9	.002
41	5	.002
42	1	.002
43	2	.001
44	14	.004
45	7	.004

Table 5.5-Cluster membership for Area

Cluster Membership		
Case Number	Cluster	Distance
1	3	482.281
2	2	2642.500
3	3	1969.281
4	4	.000
5	5	.000
6	6	914.792
7	7	381.417
8	8	605.896
9	9	109.938
10	10	.000
11	9	109.938
12	12	.000
13	15	960.406
14	14	952.000
15	15	1301.969

Cluster Membership		
Case Number	Cluster	Distance
16	3	1168.906
17	2	194.125
18	15	1167.531
19	14	1211.875
20	14	141.375
21	3	1260.469
22	8	283.021
23	3	783.719
24	11	948.458
25	7	565.792
26	14	451.125
27	1	1280.688
28	6	846.792
29	3	1443.969
30	3	693.844

Cluster Membership		
Case Number	Cluster	Distance
31	8	1026.271
32	7	947.208
33	15	825.969
34	1	1280.688
35	13	.000
36	11	477.083
37	8	710.604
38	2	804.875
39	2	2031.750
40	6	1761.583
41	8	1149.354
42	11	1425.542
43	14	49.875
44	3	561.531
45	8	55.229

Table 5.6-Cluster membership for Rectangularity

Cluster Membership		
Case Number	Cluster	Distance
1	5	.008
2	2	.002
3	3	.004
4	4	.001
5	4	.001
6	3	.005
7	7	.000
8	1	.005
9	9	.004
10	10	.000
11	11	.000
12	12	.000
13	13	.006
14	14	.006
15	14	.001

Cluster Membership		
Case Number	Cluster	Distance
16	3	.015
17	3	.011
18	13	.014
19	5	.011
20	5	.014
21	3	.012
22	3	.004
23	5	.002
24	8	.006
25	8	.002
26	6	.001
27	9	.004
28	5	.003
29	6	.001
30	3	.013

Cluster Membership		
Case Number	Cluster	Distance
31	8	.003
32	1	.007
33	13	.008
34	14	.005
35	1	.002
36	8	.005
37	6	.001
38	5	.011
39	5	.016
40	2	.002
41	15	.007
42	6	.015
43	14	.000
44	15	.007
45	6	.012

Table 5.7-Cluster membership for Diameter

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	4.500
3	3	.000
4	2	4.500
5	5	.000
6	6	.308
7	11	1.501
8	8	3.002
9	9	.000
10	7	6.004
11	11	6.004
12	11	4.503
13	13	.429
14	14	2.502
15	15	2.627

Cluster Membership		
Case Number	Cluster	Distance
16	7	1.501
17	7	3.002
18	8	3.002
19	14	2.001
20	11	3.002
21	15	2.627
22	7	7.505
23	6	1.457
24	4	4.503
25	15	1.876
26	6	3.046
27	13	4.074
28	6	7.549
29	13	2.573
30	6	6.048

Cluster Membership		
Case Number	Cluster	Distance
31	6	5.960
32	12	3.002
33	12	3.002
34	13	4.074
35	13	5.575
36	15	3.377
37	14	.500
38	13	7.934
39	10	.000
40	4	4.503
41	7	3.002
42	6	2.958
43	6	5.960
44	13	7.934
45	11	2.000E-007

Table 5.8-Cluster membership for Compactness

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	.009
3	3	.000
4	2	.009
5	5	.000
6	12	.011
7	7	.013
8	9	.020
9	9	.016
10	6	.021
11	9	.008
12	9	.009
13	13	.010
14	4	.011
15	15	.009

Cluster Membership		
Case Number	Cluster	Distance
16	12	.007
17	8	.000
18	9	.004
19	4	.004
20	7	.008
21	4	.016
22	12	.004
23	10	.001
24	14	.000
25	15	.006
26	6	8.000E-006
27	13	.006
28	10	8.050E-005
29	15	.016
30	6	.017

Cluster Membership		
Case Number	Cluster	Distance
31	6	.016
32	13	.010
33	7	.009
34	13	.001
35	13	.007
36	15	.006
37	4	.001
38	15	.002
39	9	.017
40	11	.000
41	10	.009
42	6	.011
43	10	.011
44	15	.009
45	7	.003

Table 5.9-Cluster membership for Perimeter Ratio of Major Axis-Minor Axis

Cluster Membership		
Case Number	Cluster	Distance
1	1	.034
2	2	.000
3	3	.000
4	5	.020
5	5	.020
6	6	.025
7	7	.048
8	8	.027
9	9	.015
10	9	.059
11	11	.023
12	12	.000
13	1	.033
14	14	.018
15	15	.030

Cluster Membership		
Case Number	Cluster	Distance
16	7	.057
17	14	.014
18	11	.023
19	7	.032
20	9	.044
21	7	.062
22	1	.001
23	14	.043
24	7	.039
25	14	.024
26	13	.050
27	7	.063
28	4	.018
29	13	.065
30	13	.012

Cluster Membership		
Case Number	Cluster	Distance
31	14	.041
32	13	.046
33	4	.018
34	8	.008
35	14	.027
36	13	.047
37	15	.028
38	8	.018
39	8	.046
40	6	.025
41	15	.058
42	10	.000
43	14	.029
44	13	.004
45	8	.007

Table 5.10-Cluster membership for Perimeter Ratio of Diameter

Cluster Membership		
Case Number	Cluster	Distance
1	3	.004
2	2	.000
3	3	.004
4	4	.000
5	5	.000
6	14	.087
7	1	.142
8	8	.013
9	8	.083
10	10	.059
11	11	.009
12	11	.009
13	14	.161
14	14	.007
15	14	.148

Cluster Membership		
Case Number	Cluster	Distance
16	1	.054
17	6	.026
18	15	.022
19	9	.007
20	15	.063
21	7	.072
22	10	.008
23	1	.120
24	13	.001
25	6	.024
26	10	.157
27	6	.078
28	9	.034
29	6	.079
30	6	.074

Cluster Membership		
Case Number	Cluster	Distance
31	10	.073
32	14	.039
33	6	.106
34	9	.027
35	10	.145
36	6	.017
37	10	.128
38	7	.072
39	15	.086
40	13	.048
41	12	.000
42	13	.047
43	14	.028
44	1	.076
45	8	.096

Table 5.11-Cluster membership for Concavity

Cluster Membership		
Case Number	Cluster	Distance
1	3	482.281
2	2	2642.500
3	3	1969.281
4	4	.000
5	5	.000
6	6	914.792
7	7	381.417
8	8	605.896
9	9	109.938
10	10	.000
11	9	109.938
12	12	.000
13	15	960.406
14	14	952.000
15	15	1301.969

Cluster Membership		
Case Number	Cluster	Distance
16	3	1168.906
17	2	194.125
18	15	1167.531
19	14	1211.875
20	14	141.375
21	3	1260.469
22	8	283.021
23	3	783.719
24	11	948.458
25	7	565.792
26	14	451.125
27	1	1280.688
28	6	846.792
29	3	1443.969
30	3	693.844

Cluster Membership		
Case Number	Cluster	Distance
31	8	1026.271
32	7	947.208
33	15	825.969
34	1	1280.688
35	13	.000
36	11	477.083
37	8	710.604
38	2	804.875
39	2	2031.750
40	6	1761.583
41	8	1149.354
42	11	1425.542
43	14	49.875
44	3	561.531
45	8	55.229

Table 5.12-Cluster membership for R-Factor

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	.000
3	3	.000
4	4	.000
5	5	.000
6	6	3.420
7	11	7.074
8	9	1.591
9	9	9.163
10	10	1.367
11	11	12.022
12	11	.503
13	13	.602
14	14	3.522
15	15	3.454

Cluster Membership		
Case Number	Cluster	Distance
16	10	8.599
17	7	3.755
18	9	7.571
19	14	2.811
20	11	2.049
21	15	3.454
22	7	3.755
23	6	5.167
24	12	8.423
25	15	2.458
26	6	1.634
27	13	5.083
28	6	8.314
29	13	3.178
30	6	6.101

Cluster Membership		
Case Number	Cluster	Distance
31	10	8.069
32	11	17.069
33	11	9.988
34	13	5.083
35	13	6.999
36	15	4.450
37	14	.712
38	13	9.871
39	8	.000
40	12	8.423
41	10	6.173
42	6	7.462
43	10	8.069
44	13	9.871
45	11	5.408

Table 6-Cluster membership of fruit clusters build from 14 individual fruit features

Table 6.1-Cluster membership for Branch Length

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	1.501
3	3	.000
4	4	6.380
5	4	2.627
6	6	3.002
7	9	.563
8	8	1.501
9	5	2.252
10	8	1.501
11	11	2.502
12	12	3.002
13	5	2.252
14	9	3.565

Cluster Membership		
Case Number	Cluster	Distance
15	9	5.066
16	11	2.502
17	8	1.501
18	10	.000
19	11	5.004
20	5	3.753
21	13	3.002
22	9	.938
23	14	1.501
24	2	4.503
25	5	3.753
26	8	1.501
27	9	.563
28	14	4.503

Cluster Membership		
Case Number	Cluster	Distance
29	6	4.503
30	6	.000
31	14	1.501
32	4	.375
33	14	4.503
34	13	3.002
35	9	3.940
36	7	.000
37	2	6.004
38	4	4.128
39	6	1.501
40	9	2.439
41	9	2.439
42	12	3.002

Table 6.2-Cluster membership for Branch Width

Cluster Membership		
Case Number	Cluster	Distance
1	8	5.629
2	2	4.003
3	3	3.753
4	4	4.503
5	5	3.377
6	14	4.503
7	8	2.627
8	8	.375
9	9	1.501
10	7	.300
11	11	1.201
12	9	4.503
13	2	3.502
14	14	1.501

Cluster Membership		
Case Number	Cluster	Distance
15	2	.500
16	7	4.803
17	7	7.205
18	1	5.254
19	1	5.254
20	12	.375
21	5	7.881
22	3	3.753
23	8	7.881
24	9	.000
25	5	5.629
26	11	.300
27	12	4.128
28	12	.375

Cluster Membership		
Case Number	Cluster	Distance
29	12	3.377
30	9	3.002
31	7	1.801
32	11	4.803
33	11	.300
34	11	4.203
35	5	5.629
36	7	.300
37	10	.000
38	13	.000
39	6	.000
40	4	1.501
41	4	3.002
42	14	3.002

Table 6.3-Cluster membership for Length Width Ratio

Cluster Membership		
Case Number	Cluster	Distance
1	3	.023
2	2	.002
3	3	.011
4	4	.002
5	2	.031
6	6	.000
7	7	.032
8	7	.015
9	7	.010
10	11	.021
11	11	.030
12	11	.023
13	13	.008
14	14	.021

Cluster Membership		
Case Number	Cluster	Distance
15	12	.027
16	7	.034
17	11	.002
18	11	.028
19	8	.009
20	12	.004
21	8	.015
22	7	.011
23	11	.016
24	8	.018
25	5	.015
26	8	.023
27	12	.031
28	14	.021

Cluster Membership		
Case Number	Cluster	Distance
29	4	.005
30	5	.010
31	13	.024
32	13	.016
33	13	.017
34	7	.007
35	5	.005
36	3	.013
37	1	.000
38	9	.000
39	10	.000
40	2	.021
41	2	.012
42	4	.004

Table 6.4-Cluster membership for Area

Cluster Membership		
Case Number	Cluster	Distance
1	4	2646.583
2	8	2053.063
3	2	1081.175
4	9	2388.100
5	8	3303.813
6	6	3369.958
7	4	3775.417
8	8	1675.313
9	9	899.725
10	8	3736.813
11	11	1237.800
12	8	2110.063
13	2	3292.325
14	8	2941.813

Cluster Membership		
Case Number	Cluster	Distance
15	2	420.825
16	14	2533.313
17	14	2533.313
18	4	1128.833
19	9	3817.650
20	11	3268.700
21	11	4254.175
22	12	.000
23	10	.000
24	3	909.208
25	8	2060.188
26	13	331.438
27	8	420.313
28	1	649.813

Cluster Membership		
Case Number	Cluster	Distance
29	1	649.813
30	2	932.425
31	3	1662.792
32	3	753.583
33	7	.000
34	13	331.438
35	2	1699.550
36	11	4201.575
37	5	.000
38	6	588.542
39	6	2781.417
40	11	1978.300
41	9	529.025
42	9	1058.850

Table 6.5-Cluster membership for Perimeter

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	8	34.500
3	3	9.000
4	9	25.750
5	5	.000
6	12	49.556
7	7	21.625
8	12	3.444
9	9	39.250
10	10	4.667
11	9	44.750
12	12	34.556
13	9	31.250
14	3	9.000

Cluster Membership		
Case Number	Cluster	Distance
15	7	5.375
16	12	34.556
17	6	.000
18	14	7.500
19	2	31.000
20	13	15.000
21	13	15.000
22	7	36.375
23	11	20.500
24	8	37.500
25	14	7.500
26	7	3.625
27	12	40.444
28	7	13.625

Cluster Membership		
Case Number	Cluster	Distance
29	7	20.625
30	10	21.667
31	12	2.444
32	12	39.444
33	7	11.375
34	11	20.500
35	4	.000
36	12	10.444
37	7	6.375
38	10	26.333
39	12	22.444
40	8	30.500
41	8	33.500
42	2	31.000

Table 6.6-Cluster membership for Equivalent Diameter

Cluster Membership		
Case Number	Cluster	Distance
1	4	3.052
2	8	2.413
3	2	1.340
4	9	2.705
5	8	3.882
6	6	3.687
7	4	4.361
8	8	1.967
9	9	1.015
10	8	4.387
11	11	1.366
12	8	2.485
13	2	4.083
14	8	3.465

Cluster Membership		
Case Number	Cluster	Distance
15	2	.516
16	14	3.078
17	14	3.078
18	4	1.309
19	9	4.311
20	11	3.633
21	11	4.737
22	12	.000
23	10	.000
24	3	1.183
25	8	2.422
26	13	.466
27	8	.487
28	1	.877

Cluster Membership		
Case Number	Cluster	Distance
29	1	.877
30	2	1.157
31	3	2.165
32	3	.981
33	7	.000
34	13	.466
35	2	2.102
36	11	4.663
37	5	.000
38	6	.649
39	6	3.038
40	11	2.192
41	9	.604
42	9	1.195

Table 6.7-Cluster membership for Rectangularity

Cluster Membership		
Case Number	Cluster	Distance
1	2	.006
2	2	.015
3	10	.007
4	6	.040
5	9	.009
6	5	.043
7	12	.004
8	7	.017
9	9	.016
10	10	.012
11	6	.023
12	12	.004
13	2	.007
14	5	.027

Cluster Membership		
Case Number	Cluster	Distance
15	2	.024
16	10	.005
17	7	.017
18	6	.024
19	6	.002
20	5	.022
21	5	.008
22	14	.008
23	13	.000
24	1	.003
25	2	.022
26	11	.006
27	6	.004
28	3	.014

Cluster Membership		
Case Number	Cluster	Distance
29	3	.014
30	1	.003
31	11	.006
32	8	.000
33	14	.001
34	14	.009
35	9	.007
36	5	.018
37	4	.013
38	4	.013
39	5	.016
40	5	.022
41	5	.002
42	5	.003

Table 6.8-Cluster membership for Diameter

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	3	2.252
3	13	1.501
4	10	10.695
5	5	6.004
6	6	1.876
7	9	6.380
8	7	3.002
9	9	.375
10	5	4.503
11	9	2.627
12	5	3.002
13	5	10.507
14	10	6.192

Cluster Membership		
Case Number	Cluster	Distance
15	10	2.814
16	5	4.503
17	10	1.313
18	13	1.501
19	3	2.252
20	11	.000
21	2	.000
22	4	2.502
23	10	5.817
24	8	.000
25	4	5.504
26	7	3.002
27	4	8.006
28	6	1.876

Cluster Membership		
Case Number	Cluster	Distance
29	12	3.002
30	12	3.002
31	6	4.878
32	10	.188
33	10	5.817
34	9	4.128
35	5	9.006
36	14	.000
37	5	5.000E-007
38	10	1.313
39	6	8.631
40	13	1.501
41	13	1.501
42	5	10.507

Table 6.9-Cluster membership for Perimeter Ratio of Branch Length-Branch Width

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	.039
3	3	.010
4	10	.014
5	13	.040
6	13	.031
7	7	.023
8	7	.009
9	7	.008
10	10	.024
11	10	.010
12	6	.020
13	13	.013
14	14	.000

Cluster Membership		
Case Number	Cluster	Distance
15	13	.031
16	13	.005
17	11	.024
18	5	.067
19	7	.029
20	6	.034
21	6	.000
22	7	.012
23	12	.000
24	4	.000
25	9	.007
26	3	.032
27	9	.083
28	3	.014

Cluster Membership		
Case Number	Cluster	Distance
29	6	.033
30	2	.005
31	2	.032
32	7	.002
33	2	.012
34	8	.000
35	5	.009
36	5	.058
37	9	.076
38	11	.024
39	13	.022
40	6	.021
41	3	.017
42	3	.053

Table 6.10-Cluster membership for Perimeter Ratio of Diameter

Cluster Membership		
Case Number	Cluster	Distance
1	1	.000
2	2	.044
3	3	.045
4	6	.011
5	2	.041
6	6	.091
7	11	.015
8	2	.066
9	9	.016
10	9	.065
11	11	.015
12	4	.049
13	6	.081
14	13	.080

Cluster Membership		
Case Number	Cluster	Distance
15	6	.036
16	4	.049
17	5	.000
18	2	.067
19	2	.014
20	2	.057
21	7	.037
22	2	.010
23	8	.000
24	10	.014
25	5	.000
26	9	.050
27	7	.037
28	13	.095

Cluster Membership		
Case Number	Cluster	Distance
29	13	.080
30	6	.052
31	6	.049
32	6	.119
33	6	.080
34	12	.000
35	2	.117
36	14	.000
37	3	.053
38	3	.098
39	13	.089
40	13	.083
41	13	.060
42	10	.014

Table 6.11-Cluster membership for Convexity

Cluster Membership		
Case Number	Cluster	Distance
1	1	.529
2	2	.029
3	14	.136
4	4	.370
5	5	.426
6	12	.643
7	7	.295
8	12	.045
9	4	.563
10	10	.057
11	4	.639
12	12	.450
13	4	.447
14	14	.136

Cluster Membership		
Case Number	Cluster	Distance
15	7	.073
16	12	.450
17	6	.000
18	11	.090
19	3	.505
20	8	.270
21	8	.270
22	7	.500
23	13	.000
24	9	.055
25	11	.090
26	7	.050
27	12	.515
28	7	.186

Cluster Membership		
Case Number	Cluster	Distance
29	7	.281
30	10	.268
31	12	.032
32	12	.520
33	7	.155
34	1	.529
35	5	.426
36	12	.137
37	7	.086
38	10	.325
39	12	.295
40	9	.055
41	2	.029
42	3	.505

Table 6.12-Cluster membership for Solidity

Cluster Membership		
Case Number	Cluster	Distance
1	4	.007
2	8	.005
3	2	.003
4	9	.006
5	8	.009
6	6	.009
7	4	.010
8	8	.005
9	9	.002
10	8	.010
11	11	.003
12	8	.006
13	2	.009
14	8	.008

Cluster Membership		
Case Number	Cluster	Distance
15	2	.001
16	14	.007
17	14	.007
18	4	.003
19	9	.010
20	11	.009
21	11	.011
22	12	.000
23	10	.000
24	3	.003
25	8	.006
26	13	.000
27	8	.001
28	1	.002

Cluster Membership		
Case Number	Cluster	Distance
29	1	.002
30	2	.002
31	3	.004
32	3	.002
33	7	.000
34	3	.001
35	2	.005
36	11	.011
37	5	.000
38	6	.002
39	6	.007
40	11	.005
41	9	.001
42	9	.003

Table 6.13-Cluster membership for On Pixels

Cluster Membership		
Case Number	Cluster	Distance
1	4	2664.667
2	8	2026.250
3	2	1087.200
4	9	2399.800
5	8	3284.750
6	6	3376.333
7	4	3777.333
8	8	1670.250
9	9	902.800
10	8	3740.750
11	11	1240.000
12	8	2109.750
13	2	3286.800
14	8	2930.250

Cluster Membership		
Case Number	Cluster	Distance
15	2	418.800
16	14	2549.500
17	14	2549.500
18	4	1112.667
19	9	3831.200
20	11	3284.000
21	11	4251.000
22	12	.000
23	10	.000
24	3	927.333
25	8	2075.250
26	13	316.000
27	8	433.250
28	1	649.500

Cluster Membership		
Case Number	Cluster	Distance
29	1	649.500
30	2	931.200
31	3	1671.667
32	3	744.333
33	7	.000
34	13	316.000
35	2	1687.200
36	11	4187.000
37	5	.000
38	6	585.667
39	6	2790.667
40	11	1980.000
41	9	530.200
42	9	1058.800

Table 6.14-Cluster membership for Narrow-Factor

Cluster Membership		
Case Number	Cluster	Distance
1	2	.005
2	2	.005
3	3	.000
4	7	.003
5	5	.000
6	6	.000
7	7	.001
8	6	.001
9	9	.001
10	10	.000
11	11	.000
12	11	.001
13	8	.002
14	14	.005

Cluster Membership		
Case Number	Cluster	Distance
15	14	.008
16	8	.005
17	14	.003
18	8	.007
19	4	.001
20	1	.000
21	4	.001
22	12	.000
23	7	.003
24	6	.001
25	7	.001
26	7	.001
27	5	.000
28	7	.001

Cluster Membership		
Case Number	Cluster	Distance
29	7	.001
30	9	.001
31	7	.001
32	7	.004
33	7	.001
34	13	.006
35	13	.006
36	7	.001
37	7	.001
38	7	.001
39	7	.001
40	11	.001
41	11	.005
42	11	.003

After we obtain the clusters, it is to be judged that which feature bears the best cluster formation capability. Feature having the best cluster formation capability is the most important feature for the cluster analysis. K-Means clustering algorithm does not give this feature importance measure. Hence we discuss how this measure is obtained and compare in our experiment -

We have used total 15 different classes of tomato leaves and 14 classes of tomato fruits, with each class consisting of 3 cases (patterns). Now it is to be seen that how many cases of a particular class are included in a same cluster. More is the number of cases of a particular class included in same cluster, better is clustering result. We calculate a measure called "Same Cluster Membership Ratio (SCMR)" for each class from the cluster membership (Table 5 and Table 6) produced from a particular feature and then finding the summation of SCMRs (Total SCMR) of all classes. Higher is the value of Total SCMR of a feature, higher is its importance in cluster formation. SCMR for a particular class can be calculated by finding the ratio of the total number of cases of that class included in same cluster and total number of cases present in that class (in our experiment, this value is 3 for all classes) .

Now from the above definition of SCMR, it is obvious that in our problem, we will obtain one of the following three SCMR values for each class under the following conditions-

i) If no case belonging to a particular class is included in same cluster, then SCMR of that class is 0.
ii) If 2 cases of a class are included in same cluster, then SCMR of that class is 2/3=0.67(approx).
iii) If all 3 cases of a class are included in same cluster, then SCMR of that class is 3/3=1.
Thus, the range of values of SCMR in our problem is 0 to 1.
Table 7 and Table 8 show the class number and the corresponding three case numbers (pattern number) that belong to that particular class. Cluster number is the number of the cluster; the case pattern is a member of. And the last field of Table 7 and table 8 is the calculated SCMR value of each class for each feature variables of tomato leaf and fruit. The last row of each table shows the Total SCMR value related to each leaf/fruit feature---

Table 7-SCMR calculation for individual leaf features

SCMR calculation for Major Axis			
class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	4	
	5	5	
2	6	6	0.00
	7	7	
	8	15	
3	9	9	0.00
	10	10	
	11	7	
4	12	12	0.00
	13	15	
	14	7	
5	15	15	0.67
	16	10	
	17	15	
6	18	13	0.00
	19	15	
	20	12	
7	21	10	0.00
	22	8	
	23	10	
8	24	15	0.00
	25	15	
	26	11	
9	27	10	0.00
	28	8	
	29	7	
10	30	12	0.00
	31	11	
	32	13	
11	33	7	0.00
	34	12	
	35	7	
12	36	10	0.00
	37	10	
	38	15	
13	39	14	0.00
	40	3	
	41	12	
14	42	10	0.00
	43	8	
	44	15	
15	45	13	0.00
Total SCMR			0.67

SCMR calculation for Aspect Ratio			
class number	case number	cluster number	SCMR
	1	15	
	2	8	
1	3	8	0.67
	4	11	
	5	8	
2	6	6	0.00
	7	6	
	8	8	
3	9	6	0.67
	10	8	
	11	11	
4	12	12	0.00
	13	6	
	14	1	
5	15	15	0.00
	16	3	
	17	7	
6	18	13	0.00
	19	3	
	20	5	
7	21	7	0.00
	22	3	
	23	2	
8	24	1	0.00
	25	14	
	26	14	
9	27	11	0.67
	28	14	
	29	9	
10	30	7	0.00
	31	6	
	32	1	
11	33	15	0.00
	34	1	
	35	15	
12	36	15	0.67
	37	13	
	38	5	
13	39	10	0.00
	40	6	
	41	13	
14	42	13	0.67
	43	8	
	44	1	
15	45	4	0.00
Total SCMR			3.35

SCMR calculation for Minor Axis			
class number	case number	cluster number	SCMR
	1	3	
	2	3	
1	3	3	1.00
	4	4	
	5	4	
2	6	3	0.67
	7	7	
	8	8	
3	9	7	0.67
	10	10	
	11	11	
4	12	12	0.00
	13	13	
	14	15	
5	15	15	0.67
	16	1	
	17	9	
6	18	14	0.00
	19	6	
	20	6	
7	21	9	0.67
	22	1	
	23	4	
8	24	3	0.00
	25	14	
	26	3	
9	27	2	0.00
	28	3	
	29	4	
10	30	1	0.00
	31	5	
	32	5	
11	33	5	1.00
	34	14	
	35	13	
12	36	15	0.00
	37	4	
	38	6	
13	39	9	0.00
	40	4	
	41	3	
14	42	4	0.67
	43	7	
	44	3	
15	45	6	0.00
Total SCMR			5.35

SCMR calculation for Eccentricity			
class number	case number	cluster number	SCMR
	1	13	
	2	2	
1	3	8	0.00
	4	4	
	5	8	
2	6	6	0.00
	7	9	
	8	8	
3	9	9	0.67
	10	10	
	11	11	
4	12	12	0.00
	13	6	
	14	14	
5	15	13	0.00
	16	3	
	17	7	
6	18	5	0.00
	19	3	
	20	3	
7	21	7	0.67
	22	3	
	23	1	
8	24	14	0.00
	25	14	
	26	14	
9	27	4	0.67
	28	14	
	29	3	
10	30	7	0.00
	31	6	
	32	14	
11	33	13	0.00
	34	14	
	35	13	
12	36	13	0.67
	37	5	
	38	3	
13	39	15	0.00
	40	9	
	41	5	
14	42	1	0.00
	43	2	
	44	14	
15	45	7	0.00
Total SCMR			2.68

SCMR calculation for Area

class number	case number	cluster number	SCMR
	1	3	
	2	2	
1	3	3	0.67
	4	4	
	5	5	
2	6	6	0.00
	7	7	
	8	8	
3	9	9	0.00
	10	10	
	11	9	
4	12	12	0.00
	13	15	
	14	14	
5	15	15	0.67
	16	3	
	17	2	
6	18	15	0.00
	19	14	
	20	14	
7	21	3	0.67
	22	8	
	23	3	
8	24	11	0.00
	25	7	
	26	14	
9	27	1	0.00
	28	6	
	29	3	
10	30	3	0.67
	31	8	
	32	7	
11	33	15	0.00
	34	1	
	35	13	
12	36	11	0.00
	37	8	
	38	2	
13	39	2	0.67
	40	6	
	41	8	
14	42	11	0.00
	43	14	
	44	3	
15	45	8	0.00
Total SCMR			**3.35**

SCMR calculation for Diameter

class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	2	
	5	5	
2	6	6	0.00
	7	11	
	8	8	
3	9	9	0.00
	10	7	
	11	11	
4	12	11	0.67
	13	13	
	14	14	
5	15	15	0.00
	16	7	
	17	7	
6	18	8	0.67
	19	14	
	20	11	
7	21	15	0.00
	22	7	
	23	6	
8	24	4	0.00
	25	15	
	26	6	
9	27	13	0.00
	28	6	
	29	13	
10	30	6	0.67
	31	6	
	32	12	
11	33	12	0.67
	34	13	
	35	13	
12	36	15	0.67
	37	14	
	38	13	
13	39	10	0.00
	40	4	
	41	7	
14	42	6	0.00
	43	6	
	44	13	
15	45	11	0.00
Total SCMR			**3.35**

SCMR calculation for Rectangularity

class number	case number	cluster number	SCMR
	1	5	
	2	2	
1	3	3	0.00
	4	4	
	5	4	
2	6	3	0.67
	7	7	
	8	1	
3	9	9	0.00
	10	10	
	11	11	
4	12	12	0.00
	13	13	
	14	14	
5	15	14	0.67
	16	3	
	17	3	
6	18	13	0.67
	19	5	
	20	5	
7	21	3	0.67
	22	3	
	23	5	
8	24	8	0.00
	25	8	
	26	6	
9	27	9	0.00
	28	5	
	29	6	
10	30	3	0.00
	31	8	
	32	1	
11	33	13	0.00
	34	14	
	35	1	
12	36	8	0.00
	37	6	
	38	5	
13	39	5	0.67
	40	2	
	41	15	
14	42	6	0.00
	43	14	
	44	15	
15	45	6	0.00
Total SCMR			**3.35**

SCMR calculation for Compactness

class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	2	
	5	5	
2	6	12	0.00
	7	7	
	8	9	
3	9	9	0.67
	10	6	
	11	9	
4	12	9	0.67
	13	13	
	14	4	
5	15	15	0.00
	16	12	
	17	8	
6	18	9	0.00
	19	4	
	20	7	
7	21	4	0.67
	22	12	
	23	10	
8	24	14	0.00
	25	15	
	26	6	
9	27	13	0.00
	28	10	
	29	15	
10	30	6	0.00
	31	6	
	32	13	
11	33	7	0.00
	34	13	
	35	13	
12	36	15	0.67
	37	4	
	38	15	
13	39	9	0.00
	40	11	
	41	10	
14	42	6	0.00
	43	10	
	44	15	
15	45	7	0.00
Total SCMR			**2.68**

SCMR calculation for Perimeter Ratio of Major Axis-Minor Axis

class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	5	
	5	5	
2	6	6	0.67
	7	7	
	8	8	
3	9	9	0.00
	10	9	
	11	11	
4	12	12	0.00
	13	1	
	14	14	
5	15	15	0.00
	16	7	
	17	14	
6	18	11	0.00
	19	7	
	20	9	
7	21	7	0.67
	22	1	
	23	14	
8	24	7	0.00
	25	14	
	26	13	
9	27	7	0.00
	28	4	
	29	13	
10	30	13	0.67
	31	14	
	32	13	
11	33	4	0.00
	34	8	
	35	14	
12	36	13	0.00
	37	15	
	38	8	
13	39	8	0.67
	40	6	
	41	15	
14	42	10	0.00
	43	14	
	44	13	
15	45	8	0.00
Total SCMR			**2.68**

SCMR calculation for Concavity

class number	case number	cluster number	SCMR
	1	3	
	2	2	
1	3	3	0.67
	4	4	
	5	5	
2	6	6	0.00
	7	7	
	8	8	
3	9	9	0.00
	10	10	
	11	9	
4	12	12	0.00
	13	15	
	14	14	
5	15	15	0.67
	16	3	
	17	2	
6	18	15	0.00
	19	14	
	20	14	
7	21	3	0.67
	22	8	
	23	3	
8	24	11	0.00
	25	7	
	26	14	
9	27	1	0.00
	28	6	
	29	3	
10	30	3	0.67
	31	8	
	32	7	
11	33	15	0.00
	34	1	
	35	13	
12	36	11	0.00
	37	8	
	38	2	
13	39	2	0.67
	40	6	
	41	8	
14	42	11	0.00
	43	14	
	44	3	
15	45	8	0.00
Total SCMR			**3.35**

SCMR calculation for Perimeter Ratio of Diameter

class number	case number	cluster number	SCMR
	1	3	
	2	2	
1	3	3	0.67
	4	4	
	5	5	
2	6	14	0.00
	7	1	
	8	8	
3	9	8	0.67
	10	10	
	11	11	
4	12	11	0.67
	13	14	
	14	14	
5	15	14	1.00
	16	1	
	17	6	
6	18	15	0.00
	19	9	
	20	15	
7	21	7	0.00
	22	10	
	23	1	
8	24	13	0.00
	25	6	
	26	10	
9	27	6	0.67
	28	9	
	29	6	
10	30	6	0.67
	31	10	
	32	14	
11	33	6	0.00
	34	9	
	35	10	
12	36	6	0.00
	37	10	
	38	7	
13	39	15	0.00
	40	13	
	41	12	
14	42	13	0.67
	43	14	
	44	1	
15	45	8	0.00
Total SCMR			**5.02**

SCMR calculation for R-Factor

class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	4	
	5	5	
2	6	6	0.00
	7	11	
	8	9	
3	9	9	0.67
	10	10	
	11	11	
4	12	11	0.67
	13	13	
	14	14	
5	15	15	0.00
	16	10	
	17	7	
6	18	9	0.00
	19	14	
	20	11	
7	21	15	0.00
	22	7	
	23	6	
8	24	12	0.00
	25	15	
	26	6	
9	27	13	0.00
	28	6	
	29	13	
10	30	6	0.67
	31	10	
	32	11	
11	33	11	0.67
	34	13	
	35	13	
12	36	15	0.67
	37	14	
	38	13	
13	39	8	0.00
	40	12	
	41	10	
14	42	6	0.00
	43	10	
	44	13	
15	45	11	0.00
Total SCMR			**3.35**

Table 8-SCMR calculation for individual fruit features

SCMR calculation for Branch Length			
class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	4	
	5	4	
2	6	6	0.67
	7	9	
	8	8	
3	9	5	0.00
	10	8	
	11	11	
4	12	12	0.00
	13	5	
	14	9	
5	15	9	0.67
	16	11	
	17	8	
6	18	10	0.00
	19	11	
	20	5	
7	21	13	0.00
	22	9	
	23	14	
8	24	2	0.00
	25	5	
	26	8	
9	27	9	0.00
	28	14	
	29	6	
10	30	6	0.67
	31	14	
	32	4	
11	33	14	0.67
	34	13	
	35	9	
12	36	7	0.00
	37	2	
	38	4	
13	39	6	0.00
	40	9	
	41	9	
14	42	12	0.67
Total SCMR			3.35

SCMR calculation for Length-Width Ratio			
class number	case number	cluster number	SCMR
1	1	3	0.67
	2	2	
	3	3	
2	4	4	0.00
	5	2	
	6	6	
3	7	7	1.00
	8	7	
	9	7	
4	10	11	1.00
	11	11	
	12	11	
5	13	13	0.00
	14	14	
	15	12	
6	16	7	0.67
	17	11	
	18	11	
7	19	8	0.67
	20	12	
	21	8	
8	22	7	0.00
	23	11	
	24	8	
9	25	5	0.00
	26	8	
	27	12	
10	28	14	0.00
	29	4	
	30	5	
11	31	13	1.00
	32	13	
	33	13	
12	34	7	0.00
	35	5	
	36	3	
13	37	1	0.00
	38	9	
	39	10	
14	40	2	0.67
	41	2	
	42	4	
Total SCMR			5.68

SCMR calculation for Branch Width			
class number	case number	cluster number	SCMR
	1	8	
	2	2	
1	3	3	0.00
	4	4	
	5	5	
2	6	14	0.00
	7	8	
	8	8	
3	9	9	0.67
	10	7	
	11	11	
4	12	9	0.00
	13	2	
	14	14	
5	15	2	0.67
	16	7	
	17	7	
6	18	1	0.67
	19	1	
	20	12	
7	21	5	0.00
	22	3	
	23	8	
8	24	9	0.00
	25	5	
	26	11	
9	27	12	0.00
	28	12	
	29	12	
10	30	9	0.67
	31	7	
	32	11	
11	33	11	0.67
	34	11	
	35	5	
12	36	7	0.00
	37	10	
	38	13	
13	39	6	0.00
	40	4	
	41	4	
14	42	14	0.67
Total SCMR			4.02

SCMR calculation for Area			
class number	case number	cluster number	SCMR
	1	1	
	2	8	
1	3	3	0.00
	4	9	
	5	5	
2	6	12	0.00
	7	7	
	8	12	
3	9	9	0.00
	10	10	
	11	9	
4	12	12	0.00
	13	9	
	14	3	
5	15	7	0.00
	16	12	
	17	6	
6	18	14	0.00
	19	2	
	20	13	
7	21	13	0.67
	22	7	
	23	11	
8	24	8	0.00
	25	14	
	26	7	
9	27	12	0.00
	28	7	
	29	7	
10	30	10	0.67
	31	12	
	32	12	
11	33	7	0.67
	34	11	
	35	4	
12	36	12	0.00
	37	7	
	38	10	
13	39	12	0.00
	40	8	
	41	8	
14	42	2	0.67
Total SCMR			2.68

SCMR calculation for Perimeter			
class number	case number	cluster number	SCMR
	1	1	
	2	8	
1	3	3	0.00
	4	9	
	5	5	
2	6	12	0.00
	7	7	
	8	12	
3	9	9	0.00
	10	10	
	11	9	
4	12	12	0.00
	13	9	
	14	3	
5	15	7	0.00
	16	12	
	17	6	
6	18	14	0.00
	19	2	
	20	13	
7	21	13	0.67
	22	7	
	23	11	
8	24	8	0.00
	25	14	
	26	7	
9	27	12	0.00
	28	7	
	29	7	
10	30	10	0.67
	31	12	
	32	12	
11	33	7	0.67
	34	11	
	35	4	
12	36	12	0.00
	37	7	
	38	10	
13	39	12	0.00
	40	8	
	41	8	
14	42	2	0.67
Total SCMR			2.68

SCMR calculation for Equivalent Diameter			
class number	case number	cluster number	SCMR
	1	4	
	2	8	
1	3	2	0.00
	4	9	
	5	8	
2	6	6	0.00
	7	4	
	8	8	
3	9	9	0.00
	10	8	
	11	11	
4	12	8	0.67
	13	2	
	14	8	
5	15	2	0.67
	16	14	
	17	14	
6	18	4	0.67
	19	9	
	20	11	
7	21	11	0.67
	22	12	
	23	10	
8	24	3	0.00
	25	8	
	26	13	
9	27	8	0.67
	28	1	
	29	1	
10	30	2	0.67
	31	3	
	32	3	
11	33	7	0.67
	34	13	
	35	2	
12	36	11	0.00
	37	5	
	38	6	
13	39	6	0.67
	40	11	
	41	9	
14	42	9	0.67
Total SCMR			6.03

SCMR calculation for Rectangularity			
class number	case number	cluster number	SCMR
	1	2	
	2	2	
1	3	10	0.67
	4	6	
	5	9	
2	6	5	0.00
	7	12	
	8	7	
3	9	9	0.00
	10	10	
	11	6	
4	12	12	0.00
	13	2	
	14	5	
5	15	2	0.67
	16	10	
	17	7	
6	18	6	0.00
	19	6	
	20	5	
7	21	5	0.67
	22	14	
	23	13	
8	24	1	0.00
	25	2	
	26	11	
9	27	6	0.00
	28	3	
	29	3	
10	30	1	0.67
	31	11	
	32	8	
11	33	14	0.00
	34	14	
	35	9	
12	36	5	0.00
	37	4	
	38	4	
13	39	5	0.67
	40	5	
	41	5	
14	42	5	1.00
Total SCMR			4.35

SCMR calculation for Diameter			
class number	case number	cluster number	SCMR
	1	1	
	2	3	
1	3	13	0.00
	4	10	
	5	5	
2	6	6	0.00
	7	9	
	8	7	
3	9	9	0.67
	10	5	
	11	9	
4	12	5	0.67
	13	5	
	14	10	
5	15	10	0.67
	16	5	
	17	10	
6	18	13	0.00
	19	3	
	20	11	
7	21	2	0.00
	22	4	
	23	10	
8	24	8	0.00
	25	4	
	26	7	
9	27	4	0.67
	28	6	
	29	12	
10	30	12	0.67
	31	6	
	32	10	
11	33	10	0.67
	34	9	
	35	5	
12	36	14	0.00
	37	5	
	38	10	
13	39	6	0.00
	40	13	
	41	13	
14	42	5	0.67
Total SCMR			4.69

SCMR calculation for Perimeter Ratio of Branch Length-Branch Width			
class number	case number	clusternumber	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	10	
	5	13	
2	6	13	0.67
	7	7	
	8	7	
3	9	7	1.00
	10	10	
	11	10	
4	12	6	0.67
	13	13	
	14	14	
5	15	13	0.67
	16	13	
	17	11	
6	18	5	0.00
	19	7	
	20	6	
7	21	6	0.67
	22	7	
	23	12	
8	24	4	0.00
	25	9	
	26	3	
9	27	9	0.67
	28	3	
	29	6	
10	30	2	0.00
	31	2	
	32	7	
11	33	2	0.67
	34	8	
	35	5	
12	36	5	0.67
	37	9	
	38	11	
13	39	13	0.00
	40	6	
	41	3	
14	42	3	0.67
Total SCMR			6.36

SCMR calculation for Convexity			
class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	14	0.00
	4	4	
	5	5	
2	6	12	0.00
	7	7	
	8	12	
3	9	4	0.00
	10	10	
	11	4	
4	12	12	0.00
	13	4	
	14	14	
5	15	7	0.00
	16	12	
	17	6	
6	18	11	0.00
	19	3	
	20	8	
7	21	8	0.67
	22	7	
	23	13	
8	24	9	0.00
	25	11	
	26	7	
9	27	12	0.00
	28	7	
	29	7	
10	30	10	0.67
	31	12	
	32	12	
11	33	7	0.67
	34	1	
	35	5	
12	36	12	0.00
	37	7	
	38	10	
13	39	12	0.00
	40	9	
	41	2	
14	42	3	0.00
Total SCMR			2.01

SCMR calculation for Perimeter Ratio of Diameter			
class number	case number	cluster number	SCMR
	1	1	
	2	2	
1	3	3	0.00
	4	6	
	5	2	
2	6	6	0.67
	7	11	
	8	2	
3	9	9	0.00
	10	9	
	11	11	
4	12	4	0.00
	13	6	
	14	13	
5	15	6	0.67
	16	4	
	17	5	
6	18	2	0.00
	19	2	
	20	2	
7	21	7	0.67
	22	2	
	23	8	
8	24	10	0.00
	25	5	
	26	9	
9	27	7	0.00
	28	13	
	29	13	
10	30	6	0.67
	31	6	
	32	6	
11	33	6	1.00
	34	12	
	35	2	
12	36	14	0.00
	37	3	
	38	3	
13	39	13	0.67
	40	13	
	41	13	
14	42	10	0.67
Total SCMR			5.02

SCMR calculation for Solidity			
class number	case number	cluster number	SCMR
	1	4	
	2	8	
1	3	2	0.00
	4	9	
	5	8	
2	6	6	0.00
	7	4	
	8	8	
3	9	9	0.00
	10	8	
	11	11	
4	12	8	0.67
	13	2	
	14	8	
5	15	2	0.67
	16	14	
	17	14	
6	18	4	0.67
	19	9	
	20	11	
7	21	11	0.67
	22	12	
	23	10	
8	24	3	0.00
	25	3	
	26	13	
9	27	8	0.00
	28	1	
	29	1	
10	30	2	0.67
	31	3	
	32	3	
11	33	7	0.67
	34	3	
	35	2	
12	36	11	0.00
	37	5	
	38	6	
13	39	6	0.67
	40	11	
	41	9	
14	42	9	0.67
Total SCMR			5.36

SCMR calculation for On Pixels			
class number	case number	cluster number	SCMR
	1	4	
	2	8	
1	3	2	0.00
	4	9	
	5	8	
2	6	6	0.00
	7	4	
	8	8	
3	9	9	0.00
	10	8	
	11	11	
4	12	8	0.67
	13	2	
	14	8	
5	15	2	0.67
	16	14	
	17	14	
6	18	4	0.67
	19	9	
	20	11	
7	21	11	0.67
	22	12	
	23	10	
8	24	3	0.00
	25	8	
	26	13	
9	27	8	0.67
	28	1	
	29	1	
10	30	2	0.67
	31	3	
	32	3	
11	33	7	0.67
	34	13	
	35	2	
12	36	11	0.00
	37	5	
	38	6	
13	39	6	0.67
	40	11	
	41	9	
14	42	9	0.67
Total SCMR			6.03

SCMR calculation for Narrow-Factor			
class number	case number	cluster number	SCMR
	1	2	
	2	2	
1	3	3	0.67
	4	7	
	5	5	
2	6	6	0.00
	7	7	
	8	6	
3	9	9	0.00
	10	10	
	11	11	
4	12	11	0.67
	13	8	
	14	14	
5	15	14	0.67
	16	8	
	17	14	
6	18	8	0.67
	19	4	
	20	1	
7	21	4	0.67
	22	12	
	23	7	
8	24	6	0.00
	25	7	
	26	7	
9	27	5	0.67
	28	7	
	29	7	
10	30	9	0.67
	31	7	
	32	7	
11	33	7	1.00
	34	13	
	35	13	
12	36	7	0.67
	37	7	
	38	7	
13	39	7	1.00
	40	11	
	41	11	
14	42	11	1.00
Total SCMR			8.36

Total SCMR value of each of the leaf features (Table 9) and fruit features (Table 10) defines the fact that 'Minor Axis'(Total SCMR value 5.35) ' and 'Major Axis'(Total SCMR value 0.67) are the leaf features with highest and lowest Total SCMR value. Hence 'Minor Axis' and 'Major - Axis' are the most important and least important leaf features in terms of cluster formation respectively. Where as fruit feature having the highest and lowest importance are 'Narrow-Factor'(Total SCMR value 8.36, highest among all fruit features) and 'Convexity' (Total SCMR value 2.01, lowest among all fruit features) respectively.

Table 9- Leaf features with their Total SCMR values

FEATURE NAME	TOTAL SCMR
Major Axis	0.67
Minor Axis	5.35
Aspect Ratio	3.35
Eccentricity	2.68
Area	3.35
Rectangularity	3.35
Diameter	3.35
Compactness	2.68
Perimeter Ratio of Major Axis-Minor Axis	2.68
Perimeter Ratio of Diameter	5.02
Concavity	3.35
R-Factor	3.35

Table 10- Fruit features with their Total SCMR values

FEATURE NAME	TOTAL SCMR
Branch Length	3.35
Branch Width	4.02
Length Width Ratio	5.68
Area	2.68
Perimeter	2.68
Equivalent Diameter	6.03
Rectangularity	4.35
Diameter	4.69
Perimeter Ratio of Branch Length-Branch Width	6.36
Perimeter Ratio of Diameter	5.02
Convexity	2.01
Solidity	5.36
On Pixels	6.03
Narrow-Factor	8.36

4.2 Two-step Clustering Results

Two-step clustering algorithm has used 12 and 14 number of feature variables of tomato leaf and fruit as input and 15 & 14 final clusters of leaf and fruit are produced (Figure 1).

Model Summary

Algorithm	TwoStep
Inputs	12
Clusters	15

Model Summary

Algorithm	TwoStep
Inputs	14
Clusters	14

Figure 1: Summary of Two-step clustering

Figure 2 shows the quality of the cluster formation of tomato leaf and fruit based on Silhouette measure of cohesion and separation. Silhouette measure is obtained with in the range of -1.0 to 1.0. Silhouette measure of greater than 0.5 signifies that the cluster formation is of a good quality. So from Figure 2, it is obvious that Two-step clustering has formed good quality clusters from sample patterns of tomato leaf and fruit.

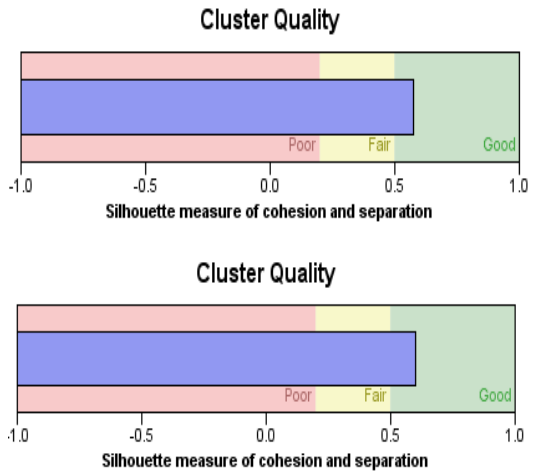
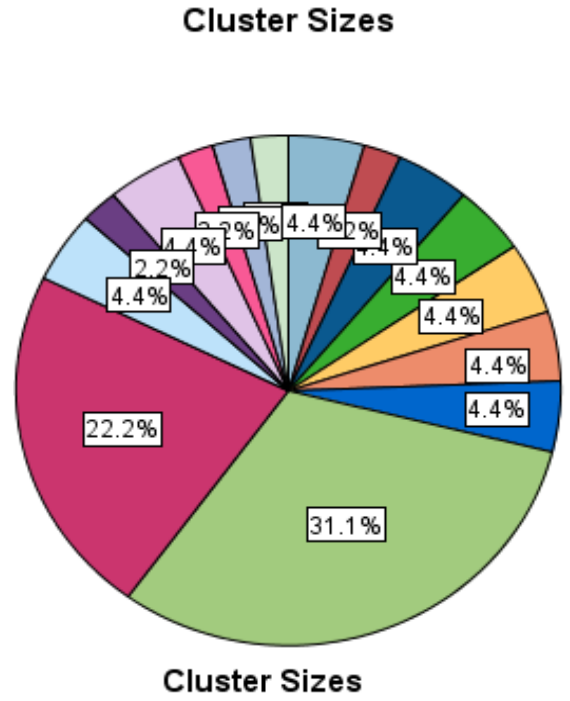


Figure 2: Cluster formation quality for leaf and fruit.

Number of patterns placed in a cluster defines the size of that cluster. Figure 3 depicts the pie chart representation of the leaf and fruit clusters according to their sizes expressed in terms of percentage of the total number of leaf and patterns.



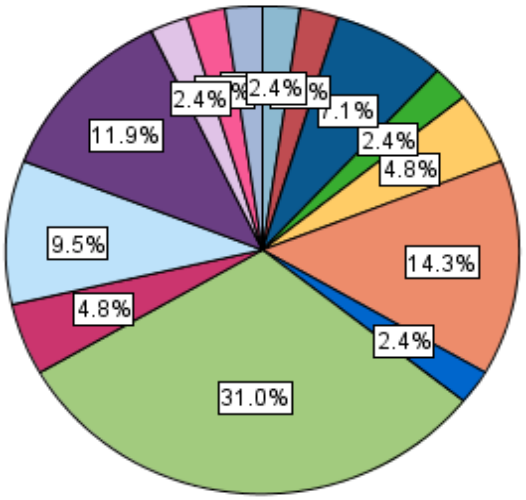


Figure 3: Pie chart of leaf and fruit clusters (size wise)

All features don't contribute in the same manner in the cluster formation process. Some features contribute highly where contribution of some features may be low. Features with higher level of contribution have higher importance in the formation of clusters. Figure 4 shows the top-down arrangement of leaf and fruit features in descending order of their importance in cluster formation. According to Figure 4, Compactness and Major Axis are the most important and least important leaf features. Where as Rectangularity and Branch Length are the fruit features with highest and lowest importance.

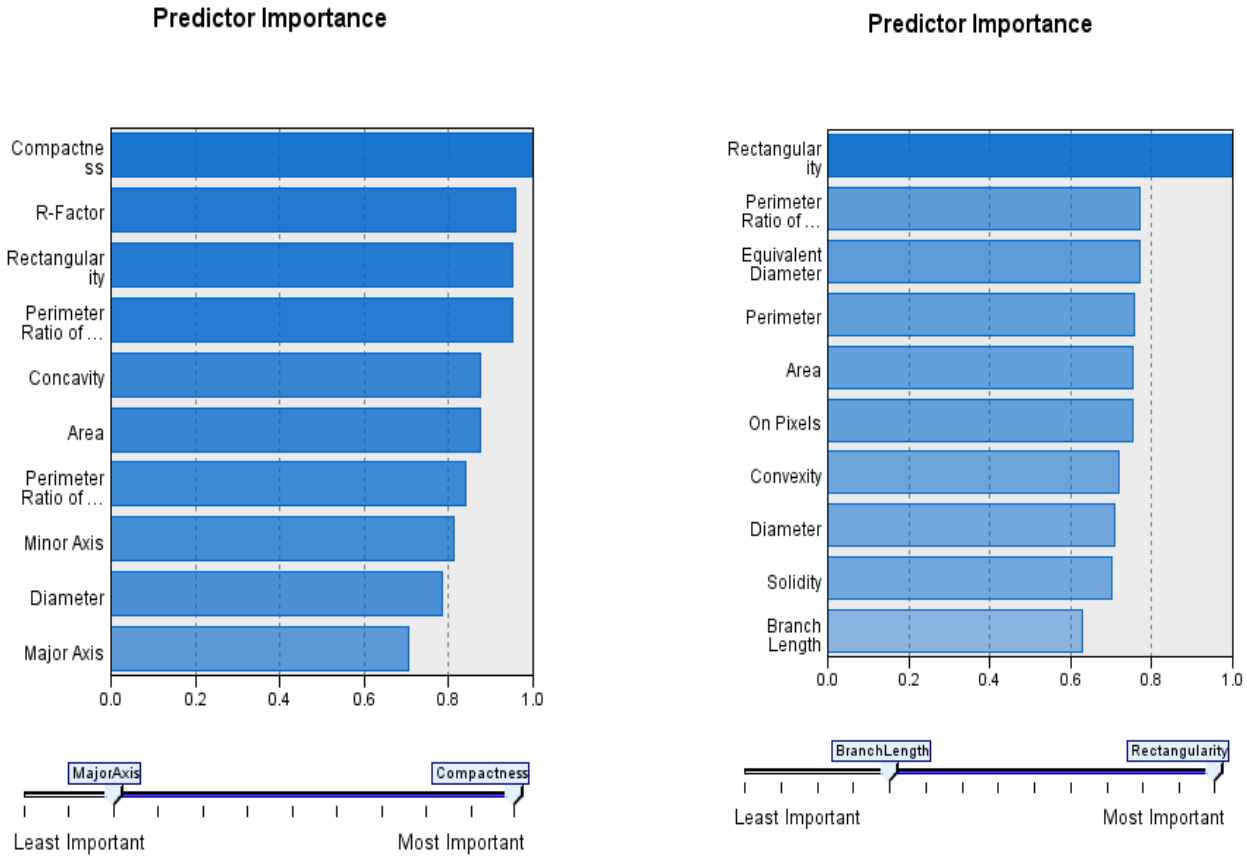


Figure 4: Order of Importance of leaf and fruit features

5. CONCLUSION

The scheme for K-Means and Two-step clustering algorithm to discriminate the tomato leaf, fruiting habit image samples along with morphological feature importance has been introduced. Formation of valid clusters assures the successful execution of the clustering techniques on the feature set and there by reflecting the categorical distribution of tomato species. Feature importance calculation through SCMR measure declares Minor Axis and Narrow-Factor as the respective leaf and fruit features upon which best cluster formation is observed. Where as according to Two-step clustering, Compactness and Rectangularity are the most important leaf and fruit feature as per as cluster formation capability is concerned. Considering this phenomena, one of the aspects of future work is to use these features to validate a large volume image dataset of tomato leaves and fruits. Another important futuristic aspect is to build up a leaf / fruit categorization system with relevant feedback mechanisms to help the persons related to cultivation process and collection of feedback from them for the enhancement of the system. Also applying the other renowned clustering methods like fuzzy clustering, neural network based clustering on the sample data set and hence analyzing the result is a future work.

6. REFERENCES

[1] A.Hazra, K.Deb, S.Kundu, P.Hazra,"Shape Oriented Feature Selection for Tomato Plant Identification", International Journal of Computer Applications Technology and Research,Volume 2-Issue 4, 2013, pages 449-454.

[2] S.Kundu, A.Hazra, K.Deb, P.Hazra, "Dimensionality Reduction of Morphological features of Tomato Leaves and Fruiting Habits", IEEE International Conference on Communications, Devices and Intelligent Systems(CODIS 2012),pages 608-611

[3] Gregory A. Wilkin, Xiuzhen Huang,"K-Means Clustering Algorithms: Implementation and Comparison", IEEE Second International Multisymposium on Computer and Computational Sciences, 2007, pages 133-136.

[4] Li-Qing Li, Qiao Liu, Han-qing Zhou,"Research on Patient Satisfaction Degree Evaluation of Three A-level Hospital in Jiangxi Province Based on Cluster Analysis", IEEE International Conference on Information Management, Innovation Management and Industrial Engineering, 2011, pages 563-567.

[5] Jie Yao, "Research on the Application of K-means cluster analysis in undergraduate instructional management", IEEE International Conference on Advanced Computer Control, 2008, pages 628-631.

[6] Bin Lie, Huichao Zhang, Huiyu Chen, Lili Liu, Dingwei Wang, "A K-means Clustering Based Algorithm for Shill Bidding Recognition in Online Auction", IEEE 24th Chinese Control and Decision Conference(CCDC 2012), pages 939-943.

[7] Shuhua Ren, Alin Fan, "*K*-means Clustering Algorithm Based On Coefficient of Variation", IEEE 4th International Congress on Image and Signal Processing, 2011, pages 2076-2079.

[8] M.Narasimha Murty, V.Susheela Devi, "Pattern Recognition An Algorithmic Approach Undergraduate Topics in Computer Science",Springer 2011, Universities Press (India) Pvt.Ltd. 207-229

[9] Andrew R. Webb, "Statistical Pattern Recognition", Second Edition, 2002, John Wiley & Sons,Ltd, 361-402

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 587 - 590, 2013, ISSN: 2319–8656

RELIABLE DATA TRANSMISSION OVER WIRELESS NETWORK USING IMAGE STEGANOGRAPHY

Roja Ramani.A
Department of Computer Science and Engineering,
TKR College of Engineering and Technology
Hyderabad, A.P-500 097, India

P.V.S. Srinivas
Department of Computer Science and Engineering,
TKR College of Engineering and Technology
Hyderabad, A.P-500 097, India

ABSTRACT Image steganography is the science of hiding data inside cover images for security. Images have a lot of visual redundancy in the sense that our eyes do not usually care about subtle changes in color in an image region. One can use this redundancy to hide text, audio or even image data inside cover images without making significant changes to the visual perception. Image steganography is becoming popular on the internet these days since a steganography image, which just looks like any other image, attracts a lot less attention than an encrypted text and a secure channel. Steganography is the science of hiding messages in such a way that no one apart from the sender and the intended recipient, suspects the existence of the message. There are multiple techniques in order to embed data in an image, however some techniques are better at being undetected then others. These techniques depend on three different aspects: capacity, security, and robustness. Capacity refers to the amount of information that can be hidden in the cover medium. Security to an eavesdropper's inability to detect hidden information. Robustness to the amount of modification the stego medium can withstand before an adversary can destroy hidden information.

KEY WORDS: Introduction, Methods to Embed files and text, LSB Algorithm LSB Embedding, Visual Attack.

---

1. INTRODUCTION

Steganography proves to be an incredibly effective way of hiding the act of communication. The ease and effectiveness of LSB embedding make it an attractive method to transmit messages without detection. With the rise in popularity of image sharing services on the Internet, it is increasingly likely that an image shared online for a short period of time would not be analyzed. It is important to note that while steganography does not guarantee that a message cannot be decoded, paring steganography with encryption provides a means of communication that is difficult to detect and can be nearly impossible for a third party to decode. While steganography can be detected by statistical attacks, relying on safety in numbers and obscure embedding patterns can limit the decoding of any particular hidden message. Steganography effectiveness, ease of implementation, and extensibility all suggest that it will be a considerable security concern for the foreseeable future. Steganography is a technique for transmitting information without detection. Steganography relies on the fact that it is difficult to detect in order to remain secure. It uses parts of an image that do not strongly influence the colours shown to embed data [3]. Where embedding is most practical varies with different image formats, but one technique that works well across formats is least significant bit embedding. Other algorithms, such as Jsteg, exploit the design of a specific image file format to embed without detection. The general principle of Steganography is that perturbing a particular value in an image using a value from the data will create a small difference in the original image. The image created by this process is a stego object. The stego object contains data from the cover and information about the data that was used to perturb the cover image. The stego object can then be decoded by the intended recipient(s) and the hidden message retrieved. Because the values in the original image are only changed slightly, an observer will struggle to visually detect that an embedding has taken place. Through this series of minor perturbations based on the message's contents the data is hidden in the cover image. A third party will have to analyze the image in order to determine if an embedding has taken place. The development of different analyses has led to an arms race between those developing steganographic algorithms and those trying to detect embeddings. From this point forward, it is assumed that the cover and data to hide are both images.

2. METHODS TO EMBED FILES AND TEXT

Image Steganography uses two methods to embed files and text into images:

1) Difference - the 'Difference' mode will output a seemingly identical image to the original input image, but this is the most noticeable mode (and thus I might remove it at a later date). By using the two images (the processed image and the original image), the 'Difference' mode compares each pixel, computes the difference, and turns it back into a byte.

2) Enlarge - the 'Enlarge' mode outputs an image 4 times bigger than the input image (2 x Width, 2 x Height). By doing this, it can has 3 times the data capacity of the 'Difference' mode, and the original image isn't required (and so this mode is the default).

3) Embed - the 'Embed' mode will output an almost identical image to the input image. It encodes the data in the last two bits of the red and Blue colour channels; but by doing this, only has half a byte per pixel.
For encryption and decryption of text messages using the secret keys steganographic system uses algorithms known as steganographic algorithms [8]. The mostly used algorithms for embedding data into images are

A. JSteg Algorithm
B. F5 Algorithm
C. LSB (Least Significant Bit) Algorithm

A. JSTEG algorithm

JSteg algorithm is one of the steganographic techniques for embedding .The hiding process will be done by replacing Least Significant Bits(LSB). JSteg algorithm replaces LSBs of quantized Discrete Courier Transform(DCT) coefficients. In this process the hiding mechanism skips all coefficients with the values of 0 or 1. This algorithm is resistant to visual attacks and offers an admirable capacity for steganographic messages[6]. Generally, JSteg steganographic algorithm embedded the messages in lossy compressed JPEG images. It has high capacity and had a compression ratio of 12%. JSteg algorithm is restricted for visual attacks and it is less immune for statistical attacks. Normally, JSteg embeds only in JPEG images. In these JPEG images, the content of the image is transformed into "frequency coefficients so as to achieve storage in
a very compressed format. There is no visual attack in the sense presented here, due to the influence of one steganographic bit up to 256 pixels[11].

B. F5 algorithm

F5 algorithm was introduced by German researchers Pfitzmann and Westfeld in order to avoid the security problem when embedding the data into the JPEG images. The F5 algorithm embeds the message into randomly chosen Discrete Courier Transform (DCT) coefficients. It utilizes matrix embedding which minimises the changes to be made to the length of certain message [5]. The F5 Algorithm provides high steganographic capacity, and can prevent visual attacks. F5 algorithm is also resistant to statistical attacks. This algorithm uses matrix encoding such that it reduces the number of changes needed to embed a message of certain length. This algorithm avoids the chi-square attack since it doesn't replace or exchange the bits. The resistance is high for both visual and statistical attacks. It has high embedding capacity that is greater than 13%.This algorithm supports TIFF, BMP, JPEG and GIFformats.The performance of the algorithms differs with the type of cover image or source on which the data is embedded[9].

C.LSB algorithm

LSB (Least Significant Bit) substitution is the process of adjusting the least significant bit pixels of the carrier image. It is a simple approach for embedding message in to the image. The Least Significant Bit insertion varies according to number of bits in an image. For an 8 bit image, the least significant bit i.e., the 8^{th} bit of each byte of the image is changed to the bit of secret message[10]. For 24 bit image, the colours of each component like RGB (red, green and blue) are changed. LSB is effective in using BMP images since the compression in BMP is lossless [10]. But for hiding the secret message inside an image of BMP file using LSB algorithm it requires a large image which is used as a cover. LSB substitution is also possible for GIF formats, but the problem with the GIF image is whenever the least significant bit is changed the whole colour palette will be changed. The problem can be avoided by only using the gray scale GIF images since the gray scale image contains 256 shades and the changes will be done gradually so that it will be very hard to detect. For JPEG, the direct substitution of steganographic techniques is not possible since it will use lossy compression. So it uses LSB substitution for embedding the data into images. There are many approaches available for hiding the data within an image: one of the simple least significant bit submission approaches is "Optimum Pixel Adjustment Procedure". The simple algorithm for OPA explains the procedure of hiding the sample text in an image [2].

Step1: A few least significant bits (LSB) are substituted with in data to be hidden
.Step2: The pixels are arranged in a manner of placing the hidden bits before the pixel of each cover image to minimize the errors.
Step3: Let n LSBs be substituted in each pixel.
Step4: Let d= decimal value of the pixel after the substitution. d1 = decimal value of last n bits of the pixel.d2 = decimal value of n bits hidden in that pixel.
Step5: If (d1~d2)<=(2^n)/2then no adjustment is made in that pixel.
ElseStep6: If(d1<d2)d = d –2^n.If(d1>d2)d=d+2^n.
This 'd' is converted to binary and written back to pixel.
This method of substitution is simple and easy to retrieve the data and the image quality better so that it provides good security.

The procedure for data hiding using steganographic application in this project is as follows

The sender first uses the steganographic application for encrypting the secret message.

3. SECRET FILE ENCRYPTION DECRYPTION TRANSMISSION

The sender first uses the steganographic application for encrypting the Secret message.
For this encryption, the sender uses text document in which the data is written and the image as a carrier file in which the secret message or text document to be hidden.
The sender sends the carrier file and text document to the encryption phase for data embedding, in which the text document is embedded into the image file. The procedure of encryption is discussed in the next phase.
In encryption phase, the data is embedded into carrier file which was protected with the password. Now the carrier file acts as an input for the decryption phase[3].
The image in which data is hidden i.e. the carrier file is sent to the receiver using a transmission medium. E.g. Web or e-mail.
The receiver receives the carrier file and places the image in the decryption phase.
In the decryption phase, the original text document can be revealed using the appropriate password.
The decryption phase decrypts the original text document using the least significant bit decoding and decrypts the

original message.

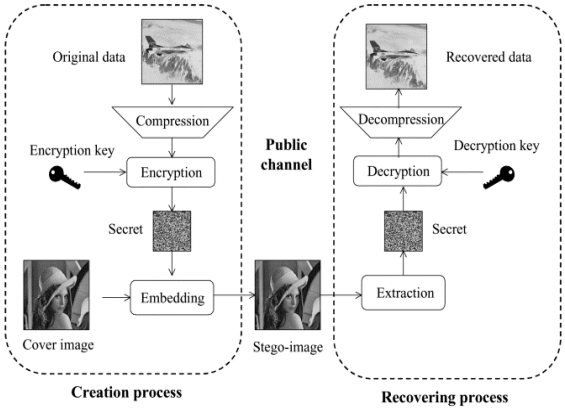


Fig-1

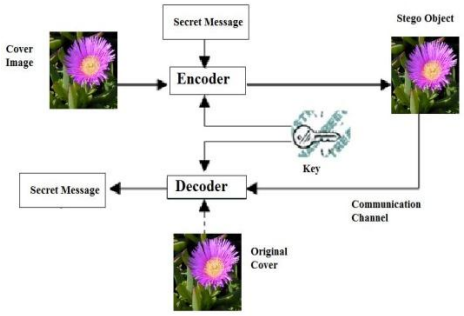


Fig-2

3.LSB Embedding

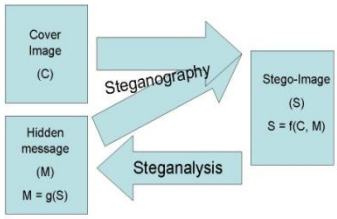


FIG -3

Least Significant Bit (LSB) embedding is a simple strategy to implement Steganography. Like all steganographic methods, it embeds the data into the cover so that it cannot be detected by a casual observer[11]. The technique works by replacing some of the information in a given pixel with information from the data in the image. While it is possible to embed data into an image on any bit-plane, LSB embedding is performed on the least significant bit(s). This minimizes the variation in colours that the embedding creates. For example, embedding into the least significant bit changes the colour value by one. Embedding into the second bit-plane can change the colour value by 2. If embedding is performed on the least significant two pixels, the result is that a colour in the cover can be any of four colours after embedding. Steganography avoids introducing as much variation as possible, to minimize the likelihood of detection. In a LSB embedding, we always lose some information from the cover image. This is an effect of embedding directly into a pixel. To do this we must discard some of the cover's information and replace it with information from the data to hide. LSB algorithms have a choice about how they embed that data to hide. They can embed losslessly, preserving all information about the data, or the data may be generalized so that it takes up less space.

Hidden files or pictures can be hidden in picture files because pictures files are so complex. Pictures on a computer are represented by tons and tons of pixels. Each pixel consists of a variation of all three primary colours, red, green and blue [1]. In a standard 24-bit bitmap, 8 bits will represent each of the three colours. 8 times 3 is 24. That means there are 256 different variations of each colour in every pixel that makes up a picture. So, to represent the colour white, the code would look like 11111111 11111111 11111111. Now, the human eye cannot distinguish the difference between too many colours and so the colour 11111110 11111110 11111110 would look exactly the same as white. Because of this, the last digit in every bit in every pixel could be changed. This is the basis of the Least Significant Bit Insertion technique. Now to show how this becomes useful. You only need 8 bits to represent Ascii text and there are three extra in every pixel of a picture. Therefore, with every three pixels, you could form one letter of Ascii text[4]. This may not seem like a lot, but when the standard image size is 640 x 480 pixels, that adds up to a lot in a hurry. In order to make this practical to the user, a computer program would be needed. After you type in your secret message and determine a cover message (the picture you want to hide you message in) the program would go through every pixel and change the last digit to represent each letter of the message you wrote. You would then send the picture to the correct recipient who would then use his program to go through every pixel and take off the last digit and use that to form the message.

The problem of using Steganography over digital communications has been solved[7]. Also, the great thing about LSB (Least Significant Bit Insertion) is that the message is not lost if the file is compressed. Anyone who uses online pictures knows that bitmap files hold a lot of information and so are generally large in size. But because the secret message is encoded into the color bits, the message is never lost when compressed. The one problem with this approach is that it does not work for every picture type. LSB works mainly with Bitmaps because of the way bitmaps are compressed[9]. JPEG's, on the other hand, are compressed using sophisticated algorithms and so a lot of the original information is lost.

Because information could so easily be lost with certain compression programs, other techniques were developed. One technique is called the Masking and Filtering technique. This technique is very similar to watermarking. The image is marked with the secret message or image and then cannot be seen unless the luminosity level is changed to an exact amount. This worked better because the text/image was now actually part of the picture and no longer in the coding part. Another technique developed used the way certain pictures are compressed to its advantage[4]. As stated earlier, JPEG's are compressed using sophisticated algorithms and because of this, a lot of the original information of the picture is lost. So, basically, what this last technique does is, it determines how the picture is going to be compressed with all the algorithms. It then changes the information of the picture accordingly to the secret message. It changes the information in a way that when decompressed, it will look similar to the LSB approach[5]. This way, when the picture is viewed, it still

looks the same but the secret message could be determined by taking the last bit of each pixel just like the LSB approach.
Today, the Internet is filled with tons of programs that uses Steganography to hide secret messages. A majority of the programs use a variation of the algorithm approach[6].

4.VISUAL ATTACK

A visual attack is the simplest way of trying to detect an embedding. It is particularly effective against LSB embeddings, but it is useless against more advanced algorithms that do not embed into the pixels of the image directly like Jsteg. A visual attack begins by looking at the image as a whole. If an embedding is detected through color abnormalities the steganographic algorithm has been successfully attacked[8]. If an embedding is not detected by the observer, the bit planes of the image are then examined, beginning with the least significant plane.

5. REFERENCES

[1]. Alfred J, M et al., 1996.Hand book of applied Cryptography.First edn.
[2]. Ali-al, H. Mohammad, A. 2010.Digital Audio Watermarking Based on the Discrete Wavelets Transform and Singular Value Decomposition,European Journal of Scientific Research. vol 39(1), pp 231-239.
[3].Amirthanjan,R. Akila,R & Deepikachowdavarapu, P., 2010. A Comparative Analysis of Image Steganography,International Journal of Computer Application , 2(3), pp.2-10.
[4]. Arnold, M. 2000. Audio watermarking: Features, applications and algorithms, Proceeding of the IEEE International Conference on Multimedia and Expo, pp 1013-1016.
[5].Bandyopadhyay, S.K., 2010.An Alternative Approach of Steganography UsingReference Image.International Journal of Advancements in Technology , 1(1), pp.05-11.
[6].Bloom,J. A. et al.,2008.Digital watermarking and Steganography . 2nd ed.
[7].MorganKaufmann.Bishop, M., 2005. Introduction to computer security. 1st ed. Pearson publications.
[8]. Cachin, C., 2004. Information: Theoretic model for steganography. Work shop on information hiding, USA.
[9]. Chan, C.K. Cheng, L.M., 2004. Hiding data in images by simple lsb substitution: pattern recognition. vol 37.
[10]. Pergamon.Cox, I. Miller, M. Bloom, J. Fridrich, J & Kalker, T. 2008. Digital watermarking and Steganography. 2ndEd. Elsevier.
[11]. Cummins, J. Diskin, P. Lau, S. & Parett, R., 2004. Steganography and digitalwatermarking. School of computer science.Vol 1.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 591 - 594, 2013, ISSN: 2319–8656

Load Adaptive Networking & Energy Aware System Design to Reduce the Energy Consumption in Telecommunication Networks

Mohd shoebuddin mujeeb
Jawahar lal Nehru Technological University
Vardhaman College Of Engineering
Shamshabad, Hyderabad, India.

S.Srinivas
Jawahar Lal Nehru Technological University
Vardhaman College of Engineering
Shamshabad, Hyderabad, India.

Abstract Worldwide, the growth rate of Internet users is about 20 percent per year. In developing countries this growth rate is closer to 40–50 percent. One of the main challenges for the future of information and communication technologies is reduction of the power consumption in telecommunication networks.
We introduce the different network architectures and the design parameters that define their power consumption. Based on these parameters the power consumption is then quantified. We elaborate on approaches for the reduction of power consumption and best communication

Keywords: Communication , Internet , Power consumption.

1. INTRODUCTION

DSL

DSL stands for Digital Subscriber Line. It is a family of technologies that provide Internet access by transmitting digital data over the wires of a local telephone network. In telecommunications marketing, bit rate of consumer DSL services typically ranges from 256 kbit/s to 40 Mbit/s downstream, depending on DSL technology, line conditions.

DSL is more like the older 'modems' in that they do use the standard copper phone lines that are in your house. It requires that there be a special switch installed at your local phone company's main routing station but DSL also has a limitation of a 4 mile.The cool feature of DSL is that you do not need to have a 'special' line installed.

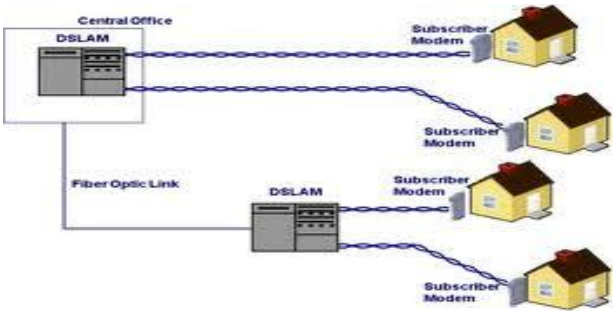


Fig. shows the Dsl System

ADSL is a type of DSL where the upstream and downstream speeds differ. There are many variations of DSL technology for different types of application (HDSL, SDSL, etc).

DSL or SDSL offers the most bandwidth but they are more expensive. In SDSL S stands for synchronous or Symmetric, this means that the bandwidth of both the upstream and the downstream are the same speed. Depending on the package you sign up for and how far away from the switch you are, the actual speed you get can vary. The most common speeds are 256k (256kilobits)/second and 1Mb (1megabit)/second. Asynchronous Digital Subscriber Line (ADSL) is same as the SDSL, but the upstream bandwidth is smaller than the downstream. Common ADSL speeds are 256kb downstream - 96kb upstream, 1mb downstream - 256kb upstream.

Peer-to-peer is a communications model in which each party has the same capabilities and either party can initiate a communication session. It will be contrasted with other models which include the master/slave model and the client/server model. In some cases, peer-to-peer communications is implemented by giving each communication node both server and client capabilities

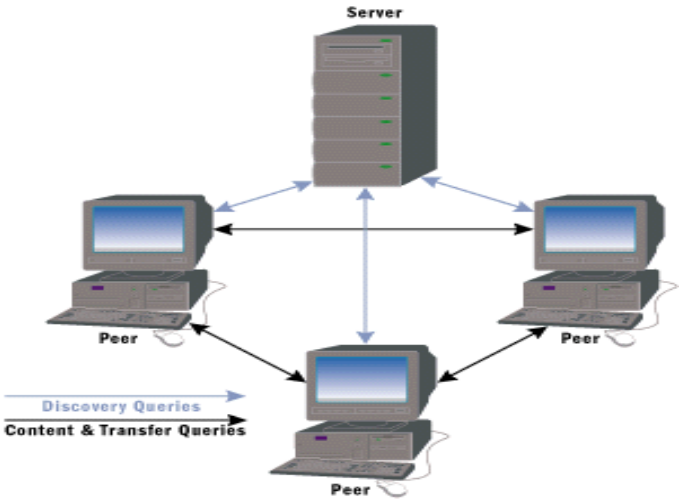


Fig. Peer to peer communication

Related work:

[1]In this paper, author demonstrated that the best opportunities for power savings come via best-of-breed technology, protocol simplification, and silicon and software optimization to achieve the least amount of processing necessary to move packets.

Pros and Cons:

--This approach is efficient under power savings.

--If a router is super-fast but lacks the required functionality and features needed to build a robust and secure communication infrastructure, it will find limited use.

[2] In this paper author proposed an anomaly detection scheme using dynamic training method in which the training data is updated at regular time intervals. Results show the effectiveness of our scheme compared with conventional scheme.

Pros and Cons:

--Presented a new detection method based on dynamically updated training data.

--Static training method could not be utilized efficiently.

[3]As Information and Communication Technology (ICT) is becoming more and more wide-spread and pervasive in our daily life, it is important to get a realistic overview of the worldwide impact of ICT on the environment in general and on energy and electricity needs in particular.

Pros and Cons:

--This mechanism decreased the power consumption.

--Power optimizations of the hardware of individual devices are not discussed.

[4] In this paper author proposed specific detection rules that can make legitimate nodes become aware of the threat, while the attack is still carrying on. Finally, the attack and present some implementation details that emphasize the little effort that an attacker would need to put in order to break into a realistic sensor network.

Pros and Cons:

--They opened the way for defining more general and formal rules in intrusion detection systems.

--Energy Efficiency of sensors is not so clarified.

[5] In this paper author examined the somewhat controversial subject motivated by data collected by the U.S. Department of Commerce i.e., energy consumption of networking devices in the Internet.

Pros and Cons:

--Energy is a precious resource whose scarcity hinders widespread Internet deployment particularly in the developing world. They showed that the impact of saving energy is huge.

--Maximization of the amount of energy conservation is not shown.

2. METHODOLOGY:

Existing system:

In existing technology, energy saving is done by putting network interfaces and other router & switch components to sleep. However, currently new approaches or variations on the suggested approaches are emerging and also the share of the core networks could become significant as well with increasing bit rates.

Is it possible to connect from one DSL modem to another (eg., from house to house)?

Yes you can connect 2 DSL users together using a VPN to make a WAN, only requirement is a router on both sides that can do the IPSec, PPTP, etc.

From your house to the CO, distance only matters for DSL. If you create a VPN then you will go from your house to the CO's router then to your friend's house but not be going directly from your house to whoever else's house, the distance only figures into if you can get DSL and at what speeds, it won't matter for VPNs.

Proposed model:

In this analysis we focus on the optical and DSL technologies. We elaborate on approaches for the reduction of power consumption. At present, in core networks the power consumption is relatively low.

Is it possible to connect from one DSL modem to another (eg., from house to house) directly?

Yes you can connect 2 DSL users together to make a communication, only requirement is a router on both sides

In our proposed model, to improve power saving there is no need of CO for providing communication, but CO is needed only for selecting best peer, communication will be done through the intermediate routers and switches…

3. MODULES:

We are dividing our project work into small modules, they are given as bellow

- Network Design
- Wireless Network Design
- Load Balancing.

To make our project as efficient we introduced one more model in our work

- Best peer selection

3.1 Network Design:

Network design and planning is an iterative process, encompassing topological design and network-realization.

a. Topological design: This stage involves determining where to place the components and how to connect them. The (topological) optimization methods

b. Network realization: This stage involves determining how to meet capacity requirements and ensure reliability within the network. This involves determining all information relating to demand and reliability and then using this information to calculate an actual physical circuit plan.

3.2 Wireless Network Design:

The energy efficiency of wireless access networks can be improved by increasing the ranges of the base stations. Thus, larger areas can be covered by a single base station, and fewer base stations are necessary. This can be done by the use of multiple transmitting and receiving antennas. This technique is known as multiple-input multiple-output.

For example, 2 transmitting and 2 receiving antennas, the range increases by 66 percent, while the power consumption increases only by 2 to 4 percent, resulting in higher energy efficiency.

3.3 Load Balancing:

Many energy aware routing algorithms and protocols have been proposed for networks recently to achieve aims like minimum energy consumption, maximized network lifetime, overhead and reduced communication latency. Load balancing is a networking method to distribute workload across multiple clusters, network links to achieve optimal resource utilization, minimize response time, maximize throughput and avoid overload.

In our proposed model, we made the model with peer to peer technology. In this model, we are using the centralized device name as server (or service provider or CO) for searching the devices, which satisfying the needs. After searching the devices we are going to make the communication through the intermediate switches without CO.

In this module, to improve load balancing we are using the two type of requesting option.

- Unicast type
- Multicast type

- Unicast
  - Mainly used in initial query to CO (it's making the problem in multiple user searches, so in that operation we are making multicasting).
- Multicast
  - Query from CO to peers (mostly multicast is used in data transmission to multiple users, but in our system we are proposing to make request).

3.4 Best Peer Selection:

In this module, we are introducing the feedback function. By using feedback analyzing requester can select best peer for communication. In feedback service user can provide positive or negative feedback. If peer is satisfying the need then user will provide '+ve' otherwise '–ve'. More +ve value shows the peer is best one.

4. RESULTS:

In this paper, we have given the model testing analysis result.

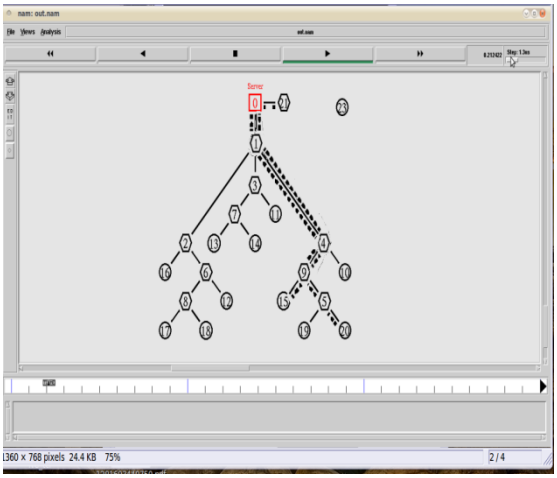


Fig. shows the model of previous system.

To improve data delivery and reduce energy consumption, we proposed a method with load balancing.

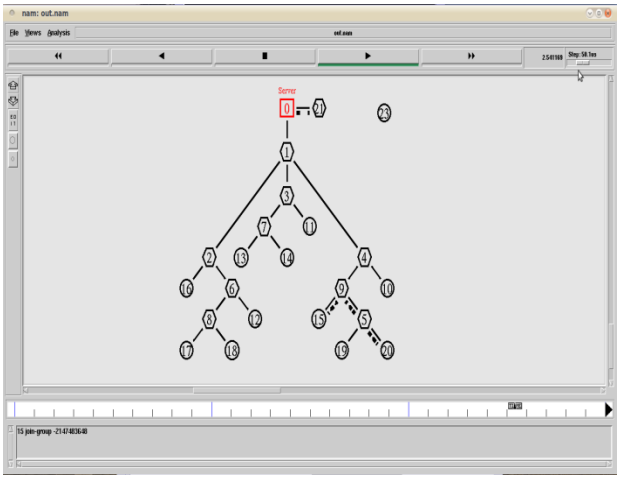


Fig. shows the model of Load Balancing system

From our analysis, we got the result with improved energy saving method. In this model we can save more energy compare than previous method. Analyzed result is given bellow.

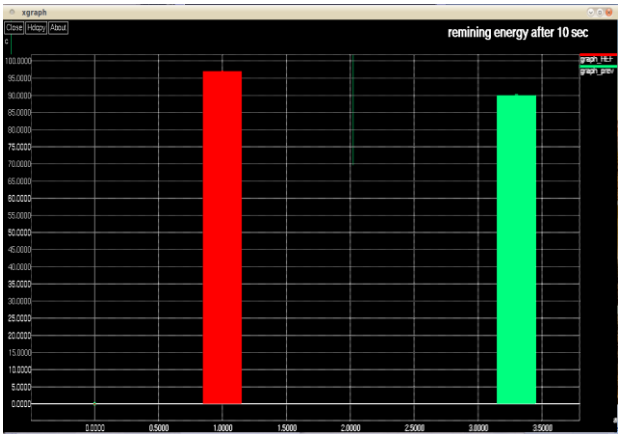


Fig. shows the Energy comparison between the Existing system and Proposed System.

First bar result is for previous model and second one is for proposed model. From this result we improved power saving upto 10%. Here we analyzed with less number of intermediate switches. If there is more intermediate switches mean we can reduce 50% of power loss.

5. CONCLUSION:

By our proposed technique, we can improve the power saving as well as high quality network performance. And also we proposed feedback based security mechanism, communication. Here we developed theory based network simulation. In future, we will implement this in real time to make green city.

6. ACKNOWLEDGEMENT:

We wish to acknowledge the efforts of ECE Dept of Vardhaman College of Engineering for guidance which helped us work hard towards producing this research work.

7. REFERENCES:

[1] "Power saving strategies and technologies in network equipment opportunities and challenges, risk and rewards"., *Luc Ceuppens, Alan Sardella, Daniel Kharitonov- 2009.*

[2] "Detecting Blackhole Attack on AODV-based Mobile Ad Hoc Networks by Dynamic Learning Method"., *Satoshi Kurosawa, Hidehisa Nakayama, Nei Kato, Abbas Jamalipour, and Yoshiaki Nemoto- 2005.*

[3] "Worldwide Energy Needs for ICT: the Rise of Power-Aware Networking"., *mario pickavet, willem vereecken, sofie demeyer, pieter audenaert – 2008.*

[4] "Launching a Sinkhole Attack in Wireless Sensor Networks; the Intruder Side"., *Ioannis Krontiris, Thanassis Giannetsos, Tassos Dimitriou – 2008.*

[5] "Greening of the Internet"., *Maruti Gupta and Suresh Singh-2003.*

[6] "Bandwidth Estimation: Metrics, Measurement Techniques, and Tools" *Ravi Prasad, Constantinos Dovrolis, Margaret Murray and kc claffy – 2003.*

[7] "Bandwidth Recycling in IEEE 802.16 Networks"., David *Chuck and J. Morris Chang – 2010.*

[8] "QoS-Oriented Intersystem Handover between IEEE 802.11b and Overlay Networks" Alexandre V. Garmonov, Seok Ho Cheon, Do Hyon Yim, Ki Tae Han, Yun Sang Park, Andrew Y. Savinkov, Stanislav A. Filin, Sergey N. Moiseev, and Mikhail S. Kondakov – 2008.

[9] "Power Consumption in Telecommunication networks: Overview and Reduction Strategies" *Willem Vereecken, Ward Van Heddeghem, Margot Deruyck, Bart Puype, Bart Lannoo – 2011.*

[10] "Dynamic Source Routing in Ad Hoc Wireless Networks"., David B.Johnson, David A. Maltz – 1996.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 595 - 600, 2013, ISSN: 2319–8656

QoS Driven Task Scheduling in Cloud Computing

Sonal Dubey
NITTTR,
Bhopal, India

Sanjay Agrawal
NITTTR,
Bhopal, India

Abstract: Cloud computing systems promise to offer pay per use, on demand computing services to users worldwide. Recently, there has been a dramatic increase in the demand for delivering services to a large number of users, so they need to offer differentiated services to users and meet their expected quality requirements. Most of scheduling schemes proceeding nowadays have no QoS (Quality of Service) differentiation, which is necessary for Cloud Computing service operation. As a cloud must provide services to many users at the same time and different users have different QoS requirements, the scheduling schemes should be developed having different QoS requirements. So, this paper explores various methods of task scheduling done in cloud computing. Real-time applications play a significant role in cloud environment. We have examine the particular scheduling algorithms for real-time tasks, that is, priority-based strategies.The purpose of this paper is to discuss the fixed priority preemptive task scheduling algorithms in cloud computing for improving the QoS parameters.

Keywords: Cloud Computing, QoS, RMS, DMS, UB Test.

1. INTRODUCTION

Cloud computing is the rising technology that delivers many forms of resources as services, mainly over the internet. It permits customers to use applications without deployment and access the required files at any computer using internet [3]. Cloud Computing allows on-demand resource provisioning. It is the convergence of several concepts such as grid, distributed application design, virtualization and enterprise IT management. It enables a more flexible approach for deploying and scaling applications [2].

The Task management is the key role in cloud computing. Task scheduling problems are primary which relate to the efficiency of the whole cloud computing facilities. Because of different QoS parameters such as CPU speed, CPU utilization, turnaround time, throughput, waiting time etc., task scheduling in cloud computing is different from conventional distributed computing environment. The demand for resources fluctuates dynamically so scheduling of resources is a difficult task. Task scheduling based on QoS parameters is necessary for efficient resource utilization and for satisfying user requirement.

Scheduling in cloud computing can be categorized into three stages namely–

- Resource discovering and filtering – Data centre Broker discovers the resources present in the network system and collects status information related to them.
- Resource selection – Target resource is selected based on certain parameters of task and resource. This is key stage.
- Task submission -Task is submitted to resource selected.

The goal of scheduling algorithms in distributed systems is to schedule jobs to the flexible resources in accordance with flexible time, which includes finding out a proper sequence in which jobs can be executed under transaction logic constraints. The main advantage of task scheduling algorithm is to achieve a high performance computing and the best system throughput.

Here, we consider the following terms for our understanding:

- Task: ti
- Virtual machine: mj
- Time when task ti arrives: ci
- Time when machine mj is available: aj
- Execution time for ti on mj: eij
- Time when the execution of ti is finished on mj, cij=aj + eij: cij
- Maximum value of cij: makespan

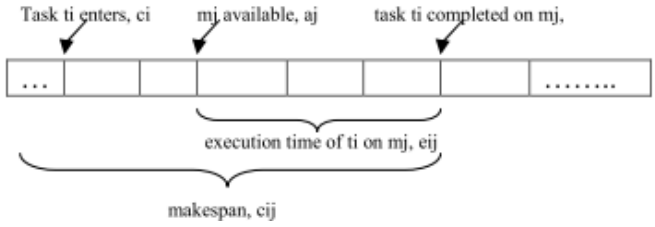


Figure 1. Task Scheduling

Cloud computing has been defined by NIST "as a model for supporting convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction". This cloud model is composed of five essential characteristics, three service models, and four deployment models. The five essential characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The three service models are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (laaS). The four deployment models are Private Cloud, Community Cloud, Public Cloud, and Hybrid Cloud [1].

A scheduling is a set of rules that determine the jobs to be executed at a particular time. This paper is concerned only with fixed-priority pre-emptive scheduling, which works as

follows. A distinct and fixed priority is assigned to each task. When a job is initiated with a priority higher than the one currently being executed, the current job is immediately interrupted and the new job is started. The remaining paper is ordered as follows. Section II describes RMS and DMS algorithms. Section III presents the related work in this area. Section IV concludes the paper.

2. FIXED PRIORITY SCHEDULING IN CLOUD COMPUTING

In real time cloud applications the cloud users and providers must have a strong service level agreement to ensure the timing of applications, and deadlines of applications. A real-time scheduler must ensure that processes meet deadlines, regardless of system load or makespan. Priority is applied to the scheduling of these periodic tasks with deadlines. Every task in priority scheduling is given a priority through some policy, so that scheduler assigns tasks to resources according to priorities.

In fixed priority scheduling the dispatcher will make sure that at any time the highest priority runnable task is essentially running. So, if we have a task with a low priority running and a high priority task arrives. The low priority task will be suspended and the high priority task will start running. If while the high priority task is running a medium priority task arrives the dispatcher will leave it unprocessed and the high priority task will carry on running and at some later time finish its computation. So the task with medium priority starts executing to finish at some later time. Only when both tasks have completed can the low priority task resume its execution. The low priority task can then carry on executing until either higher priority tasks arrive or it has finished its work. The fixed priority scheduling algorithms help in improving the QoS parameters in cloud computing environment as they have less runtime overhead, robust, optimal and easy to implement.

2.1 Related Terms

The deadline of a request for a task is defined to be the time of the next request for the same task. For a set of tasks scheduled as per some scheduling algorithm, an overflow occurs at time 't' if 't' is the deadline of an unfulfilled request. For a set of tasks, a scheduling algorithm is feasible if the tasks are scheduled so that no overflow ever occurs. We define the response time of a request for a certain task to be the time span between the request and the end of the response for the request. A critical instant is defined as an instant at which a request for that task will have the largest response time. A schedulability test is a mechanism that proofs that all deadlines are met, when scheduling with a particular algorithm. The schedulable utilization of a scheduling algorithm is defined as follows: A scheduling algorithm can feasibly schedule *any set* of periodic tasks on a processor if the total utilization of the tasks is equal to or less than the schedulable utilization of that algorithm.

2.1.1 Periodic Task Model

- A task = (C, T)

  C: worst case execution time/computing time (C<=T!)

  T: period (D=T)
- A task set: (Ci,Ti)

  All tasks are independent

The periods of tasks start simultaneously at time 0

- C/T is the CPU utilization of a task
- U = Σ (Ci/Ti) is the CPU utilization of a task set

2.2 Rate Monotonic Scheduling algorithm (RMS)

It is a dynamic pre-emptive algorithm for scheduling set of independent hard real time tasks. This was published in 1973 by Liu and Layland [5]. The algorithm was based on static task priorities. The assumptions made about the task set are mentioned below [3, 4].

1. The request for all the task sets is periodic.

2. All tasks are independent of each other. No precedence constraints or mutual exclusion constraints exist between any pair of tasks.

3. The deadline interval of every task is equal to its period.

4. The required maximum computation time is known beforehand and is constant.

5. Time required for context switching can be ignored.

6. Sum of utilization factors of n tasks with period p is given by $U=\Sigma (ci/pi) \leq n (21/n - 1)$. As n approaches infinity, term $n (21/n - 1)$ reaches ln 2 (about 0.7).

The task priorities are assigned on the basis of their periods. The task with shortest period gets the highest priority and the task with longest period gets lowest priority. If all the assumptions stated above are satisfied then this algorithms guarantees that all the tasks will meet their deadlines. The algorithm is optimal for single processor systems.

2.2.1 Basic properties of rate monotonic scheduling

For each task that is to be scheduled we must know the value of its period T and the worst case performance time C, so that the value of the processor load coefficient could be calculated as C/T.

If the set of the tasks being scheduled is given and the characteristics of the tasks are known, a significant question is whether the time constraints of all the tasks will always be met. This is answered by the Liu and Layland theorem which is given by the following formula:-

$$\sum_{i=1} (C_i / T_i) \leq n (2^{1/n} - 1) \qquad (1)$$

In the inequality (1), n is the number of the tasks scheduled. The inequality (1) delivers only a sufficient condition for the set of schedulable tasks. However, the condition (1) is not a necessary condition for the set of schedulable tasks. Furthermore, if the condition (1) is not fulfilled, it does not automatically follow that the set of tasks is not schedulable. In this case, one must check whether the necessary condition is fulfilled. The necessary condition is given by the following formula

$$\sum_{i=1}^{n} (C_i / T_i) \leq \quad (2)$$

or

n

$$_i \min \sum_{i=1} (C_i / lT_k) \lceil lT_k / T_i \rceil \leq 1 \qquad (i= 1,2\ldots..,n) \quad (2)$$

where, min is calculated over $(k,l) \in W_i$

and $W_i = \{ (k,l) , 1 \leq k \leq i, l = 1,\ldots.., \llcorner T_i / T_k \lrcorner$

Moreover, for each task of the scheduled set of tasks it needs to be checked, whether their time constraints are met in the worst case situation, i.e. under the conditions when all the tasks enter into the ready state at the same moment. If under the worst-case conditions the performance of all the scheduled tasks is ended before the elapse of their time constraints, it means that the given set of tasks is schedulable under any circumstances. In order to prove this, one has to calculate for each task the time when its execution end. If the execution end time of each task is shorter than its time deadline, it means that the set of tasks is schedulable [5].

To calculate the time of the execution end of a periodic task the recurrent formula can be used. If we consider the lowest priority task, then the first approximation of its time of the execution end is assumed as the sum of its execution time and the times of execution of all the other tasks. This results from the fact that before the execution of the lowest priority task can be started, all the other tasks must be performed at least once. Thus, the first estimation of the time of the execution end of a task is given by the following formula

$$t_{0=} \sum_{i=1}^{n} C_i \qquad (3)$$

Then, we must systematically repeat the recurrent procedure, which is given by the following formula

$$t_{m+1 =} \sum_{i=1}^{n} C_i . \lceil t_m / T_i \rceil \qquad (4)$$

The recurrent procedure is repeated until the following condition is fulfilled

$$t_{m+1 =} \; t_m \qquad (5)$$

In such a case we consider time tm as the time of the execution end of the lowest priority task. If this time is shorter than the deadline of the lowest priority task, we can consider this task schedulable under any circumstances, because it proved to be schedulable in the worst-case scenario.

The recurrent procedure, which is discussed above, must be repeated for all the tasks and all the tasks in the worst-case scenario must be proved to be able to end their executions before the elapse of their deadlines. Only if this condition is met, the given set of periodic tasks may be considered schedulable.

2.2.2 Sufficient Schedulability Test: Utilization Bound Test (UB Test)

Assume a set of n independent tasks: S= {(C1,T1) (C2,T2)...(Cn,Tn)}and Let U= Σ Ci/Ti and B(n) =n*($2^{1/n}$-1)

Three possible outcomes:

- 0<= U<= B(n): schedulable
- B(n)<U<=1: no conclusion
- 1< U : overload

2.2.3 Sufficient and Necessary Schedulability Test

1. Calculate the worst case response time R for each task with deadline D. If R<=D, the task is schedulable/feasible. Repeat the same check for all tasks
   - Ri= Ci + ∑j ∈ HP(i) ⌈ Ri/Tj ⌉*Cj

⌈ Ri/Tj ⌉ is the number of instances of task j during Rj

⌈Ri/Tj ⌉ *Cj is the time needed to execute all instances of task j released within Rj

∑j ∈ HP(i) ⌈ Ri/Tj ⌉*Cj is the time needed to execute instances of tasks with higher priorities than ith task, released during Rj

Rj is the summation of the time required for executing task instances with higher priorities than task j and its own computing time

- We need to solve the equation:

  Ri= Ci + ∑j ∈ HP(i) ⌈ Ri/Tj ⌉*Cj

- This can be performed by numerical methods to calculate the fixed point of the equation by iteration: let

Ri0 = Ci + ∑j ∈ HP(i) Cj = C1 + C2 +…..+ Ci (the first guess)

Ri k + 1 = Ci + ∑j ∈ HP(i) ⌈ Ri/Tj ⌉*Cj (the (k+1)th guess)

- The iteration stops when either

  Rim + 1 > Ti or non-schedulable

  Rim + 1 < Ti and Rim + 1 = Rim schedulable

2. If all tasks pass the test, the task set is schedulable.
3. If some tasks pass the test, they will meet their deadlines even the other don't (stable and predictable).

The following rule of thumb can be given to simplify the schedulability check by RMS:

Step1. Apply Equation (1) and stop if all individual conditions are met. If not, apply Equation (2) for all doubtful cases, as in the next Steps (Steps 2a – 2c).

Step2a. Determine all schedulability Points by marking on a time axis all successive periods for all tasks in question, from time 0 up to the end of the first period of the lowest-frequency task.

Step2b. For each time instant marked in Step 2a - that is, for all schedulability Points - construct an inequality that has, on its left-hand side, a sum of all possible execution times of all tasks that can be activated (possibly multiple times) before this schedulability Point and, on its right-hand side, only the value of time corresponding to this schedulability Point.

Step2c. Check if values on the left-hand sides are smaller than or equal to their corresponding right-hand-side values. If at

least one of these inequalities holds, then the set of tasks is schedulable according to RMS Equation (2). If not, then the set of tasks is not schedulable according to RMS.

2.3 Deadline Monotonic Scheduling algorithm (DMS)

This technique is an extension of Rate Monotonic scheduling algorithm. This is first proposed in 1982 by Leung and Whiteland. This is also fully pre-emptive technique used for scheduling tasks with static priorities [3]. The third assumption mentioned in rate monotonic technique that says the deadline interval of every task is same and equal to its period has been relaxed. The tasks have deadlines that relative deadlines (Di) can be less than or equal to its period. Each task is allotted a fixed priority inversely proportional to its relative deadline. So, at any time task with the shortest deadline is executed. Deadline monotonic is a static priority scheduling method, as relative deadlines are constant.

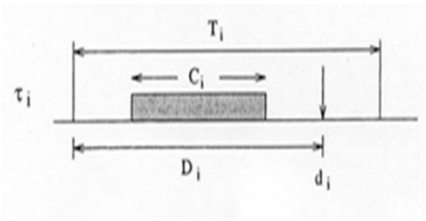


Figure2. Deadline Monotonic Scheduling

There are four parameters for each periodic task:

- A phase ϕi;
- A worst- case computation time Ci (constant for each instance);
- A relative deadline Di (constant for each instance);
- A period Ti;

which have the following relationships:

- $Ci \leq Di \leq Ti$
- $ri,k = \phi i + (k-1)Ti$
- $di,k = ri,k + Di$

2.3.1 Sufficient schedulability test: Utilization Bound test (UB test)

$\Sigma (Ci/Di) <= n*(21/n-1)$ implies schedulable by DMS

2.3.2 Precise test

- Calculate the worst case response time R as described above, for each task with deadline D. If R<=D, the task is schedulable/feasible. Repeat the same check for all tasks.
- If the test is passed by all the tasks, the task set is schedulable.
- If the test is passed by some tasks only, they will meet their deadlines even the other doesn't (stable and predictable).

3. RELATED WORK

Tradition ways of task scheduling is not fit to Cloud Computing [1][7][8]. At present, there are lots of task scheduling schemes implemented in different cloud framework. Hadoop [9] implements the FIFO (First In First Out) [10] scheme by default. The benefit of FIFO is simple and low overhead. All the jobs from different users are submitted to a queue. After that they will be examined according to the order of submission time and priority. The first job with highest priority will be selected for processing. The disadvantage of FIFO is poor fairness. The jobs with lower priority have little chance to process with lots of higher priority jobs.

In order to improve the fairness, Facebook presented Fair Scheduling Algorithm [6]. The goal of fair scheduling is that all tasks can achieve their resource as the time passes. This algorithm allows short tasks finish in reasonable time while not starving long tasks. Task occupies the whole resource with no other tasks in the system. And the system will allocate the idle time slot to those new tasks and make each of them could get equal CUP time. Fair Scheduling defines insufficiency of tasks. Tasks with more shortfalls mean they got more unfair treatments, so they have more probability to obtain resource. Apart this, fair scheduling algorithm assured the minimized shared resource. It means task with lowest priority might have its turn even if there are many tasks with higher priority.

Yahoo! presents Capacity Scheduling for Hadoop as well [11]. It allows for multiple-tenants to securely share a large cluster such that their applications are allocated resources in a timely manner under constraints of allocated capacities. This scheme allows sharing a large cluster while giving each organization a minimum capacity guarantee. Clusters will be partitioned among multiple organizations and each organization can access any excess capacity no being used by others.

All the algorithms introduced above focus on tasks of computing oriented, and not fit for service oriented tasks. In addition, Lee et al. presented a dynamic priority-scheduling algorithm on service request scheduling [12]. It adjusts the priority of task units on scheduling to increase the performance of scheduling. Yoshitomo et al. presented a history-based job scheduling mechanism for a queue system [13]. This mechanism estimates the time to start the job execution according the history of job-execution and the jobs scheduling mechanism automatically allocates the job to a suitable resource. Luqun Li offered an optimistic differentiated service task scheduling system. This paper developed a non-pre-emptive priority M/G/1 queuing model for the tasks and the system cost function for this model. Subsequent to that, the author gave the corresponding strategy and algorithm to get the approximate optimistic value of service [14]. QuXilong and Hao Zhongxiao et al. researched the distributed software resource sharing in Cloud Manufacturing system and implemented the sharing scheme in a cloud platform [15].

However, the scheduling schemes introduced above are centralized algorithms and will become bottleneck in large scale Cloud Computing environment. Moreover, they are designed for a precise computing concept, which is performance oriented and not suitable for other Cloud Computing Services, which are service oriented. The earlier one executed with short period and high utility and the later one executed with long term and lower utility [16].

- RSDC (Reliable Scheduling Distributed in Cloud computing)

Arash Ghorbannia Delavar,Mahdi Javanmard , Mehrdad Barzegar Shabestari and Marjan Khosravi Talebi [1] proposed a reliable scheduling algorithm in cloud computing environment. In this algorithm main job is divided into sub jobs. To balance the jobs, the request and acknowledge time are calculated independently. Scheduling of each job is done by calculating the request and acknowledges time in the form of a shared job, so that the efficiency of the system is increased.

- An Optimal Model for Priority based Service Scheduling Policy for Cloud Computing Environment

Dr. M. Dakshayini, Dr. H. S. Guruprasad [3] proposed a new scheduling algorithm based on admission and priority control method. In this algorithm, priority is assigned to each admitted queue. Entrance of each queue is decided by calculating tolerable delay and service cost. The advantage of this algorithm is that with the proposed cloud architecture this scheme has achieved very high (99%) service completion rate with definite QoS. As this scheduling provides the highest preference for highly paid user requests for service, total servicing cost for the cloud also increases.

- A Priority based Job Scheduling Algorithm in Cloud Computing

Shamsollah Ghanbari, Mohamed Othman [17] proposed a new scheduling algorithm based on multi – criteria and multi - decision priority driven scheduling algorithm. This scheduling algorithm have three levels of scheduling: object level, attribute level and alternate level. This algorithm set the priority by job resource ratio. Next priority vector can be compared with each queue. This algorithm has high throughput and less finish time.

- Extended Max-Min Scheduling Using Petri Net and Load Balancing

El-Sayed T. El-kenawy, Ali Ibraheem El-Desoky, Mohamed F. Al-rahamawy [5] has proposed a new algorithm based on impact of RASA algorithm. Extended Max-min algorithm is based on the expected execution time rather on complete time as a selection basis. To model the concurrent behavior of distributed systems Petri nets are used. Max-min algorithm shows achieving schedules with comparable lower makespan rather than RASA and original Max-min.

- An Optimistic Differentiated Job Scheduling System for Cloud Computing

Shalmali Ambike, Dipti Bhansali, Jaee Kshirsagar, Juhi Bansiwal [6] has proposed a differentiated scheduling algorithm with non-preemptive priority queuing model for activities performed by cloud user. In this method, a web application is created to do some activity like one of the file uploading and downloading then there is need of efficient job scheduling algorithm. This algorithm helps in achieving the QoS requirements of the cloud computing user and the maximum profits of the cloud computing service provider.

- Improved Cost-Based Algorithm for Task Scheduling

G.Mrs.S.Selvarani, Dr.G.Sudha Sadhasivam [7] proposed an improved cost-based scheduling algorithm for making efficient mapping of tasks to available computing resources in cloud environment. The managing of traditional activity based costing is proposed by new task scheduling strategy for cloud environment where there may be no relation between the overhead application base and the way that different tasks cause overhead cost of resources in cloud. The proposed algorithm divides all user tasks depending on priority of each task into three different lists. The proposed algorithm calculates both resource cost and computation performance. It also improves the computation/communication ratio.

- Performance and Cost evaluation of Gang Scheduling .in a Cloud Computing System with Job Migrations and Starvation Handling

Ioannis A. Moschakis and Helen D. Karatza proposed a gang scheduling algorithm with job migration and starvation handling. The number of Virtual Machines (VMs) available at any moment is dynamic and scales according to the demands of the jobs being serviced. The above mentioned model is studied through simulation in order to analyze the performance and overall cost of Gang Scheduling with migrations and starvation handling. Results show up that this scheduling strategy can be effectively deployed on cloud environment, and that cloud platforms can be feasible for HPC or high performance enterprise applications.

4. CONCLUSION

Scheduling is one of the most important accept in cloud computing environment. So, in this paper, we focused on task scheduling in cloud computing environment with certain QoS constraint. Rate Monotonic algorithm is simpler to implement and exhibits a predictable behaviour resulted from its associated analysis. In this paper, fixed priority scheduling algorithms i.e. Rate Monotonic and Deadline Monotonic scheduling algorithms are explained, when using them in cloud computing environment for improving the QoS parameters

REFERENCES

[1] Peter Mell, Timothy Grance, “The NIST Definition of Cloud Computing”, NIST (National Institute of Standards and Technlogy) Special Publication 800-145.

[2] Jun Huang, Yanbing Liu, Qiang Duan, “Service Provisioning in Virtualization-based Cloud Computing: Modeling and Optimization”, Globecom, 2012.

[3] Q. Duan, “Modeling and Performance Analysis on Network Virtualization for Composite Network-Cloud Service Provisioning,” in Proc. of SERVICES, 2011, pp. 548–555, July 2011.

[4] Z. Wang and J. Crowcroft, “Quality-of-service routing for supporting multimedia applications,” IEEE J. Sel. Areas. Commun., vol. 14, no. 7, pp. 1228–1234, Sept. 1996.

[5] G. Xue, A. Sen, W. Zhang, J. Tang and K. Thulasiraman, “Finding a path subject to many additive QoS constraints,” IEEE/ACM Trans. Netw., vol. 15 , no. 1, pp. 201-211, Feb. 2007.

[6] G. Xue, W. Zhang, J. Tang and K. Thulasiraman, "Polynomial time approximation algorithms for multi-constrained QoS routing," IEEE/ACM Trans. Netw., vol. 16, no. 3, pp. 656-669, Jun. 2008.

[7] J. Huang, X. Huang, and Y. Ma, "An Effective Approximation Scheme for Multi constrained Quality-of-Service Routing," in Proc. IEEE GLOBECOM 2010, Miami, Florida, pp. 1–6, Dec. 2010.

[8] F. A. Kuipers, P.V. Mieghem, T. Korkmaz and M. Krunz, "An Overview of Constraint-Based Path Selection Algorithms for QoS Routing," IEEE Com. Mag., vol. 40, no. 12, pp. 50–55, Dec. 2002.

[9] Z. Tarapata, "Selected multi criteria shortest path problems: an analysis of complexity, models and adaptation of standard algorithms," Int. J. Applied Math. and Comp. Sci., vol. 17, no. 2, pp. 269–287, July 2007.

[10] R. G. Garroppo, S. Giordano and L. Tavanti, "A survey on multi constrained optimal path computation: Exact and approximate algorithms," Compt. Netw., vol. 54, no. 17, pp. 3081–3107, Dec. 2010.

[11] X. Yuan, "Heuristic Algorithms for Multi constrained Quality-of-Service Routing," IEEE/ACM Trans. Netw., vol. 10, no. 2, pp. 244–256, Apr. 2002.

[12] P. V. Mieghem and F.A. Kuipers, "Concepts of Exact QoS Routing Algorithms," IEEE/ACM Trans. Netw., vol. 12, no. 5, pp. 851–864, Oct. 2004.

[13] Jun Huang, et al., "Novel End-to-End Quality of Service Provisioning Algorithms for Multimedia Services in Virtualization-based Future Internet", IEEE Transactions On Broadcasting.

[14] Yee Ming Chen, Yi Jen Peng, "A QoS aware services mashup model for cloud computing applications" Journal of Industrial Engineering and Management, JIEM, 2012 – 5(2): 457-47.

[15] Pinal Salot Purnima Gandhi , "Job Resource Ratio Based Priority Driven Scheduling in Cloud Computing", International Journal for Scientific Research & Development, Vol. 1, Issue 2, 2013 , ISSN (online): 2321-0613.

[16] Bing Li, A Meina Song, Junde Song, "A Distributed QoS-Constraint Task Scheduling Scheme in Cloud Computing Environment: Model and Algorithm", Advances in information Sciences and Service Sciences(AISS), Volume4, Number5, March 2012.

[17] Ghanbari Shamsollah, and Othman Mohamed, " A Priority based Job Scheduling Algorithm in Cloud Computing", Procedia Engineering, Vol. 50, pp. 778-785, 2012.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 601 - 605, 2013, ISSN: 2319–8656

A Note on the New Similarity Measure for Fuzzy sets

Pranamika Kakati
Department of Computer Application,
Girijananda Chowdhury Institute of Management & Technology,
Guwahati,Assam,India.

Abstract: In this article, we intend to draw attention on the new Similarity measure for Fuzzy sets based on the extended definition of complementation. The old existing measures are based on traditional Zadehian Theory of Fuzzy sets where it is believed that there is no difference between Fuzzy membership function and Fuzzy membership value for the complement of a Fuzzy set which is already proved to be wrong. As a result, the previous Similarity measures have been proved illogical from the standpoints of new definition of complementation of Fuzzy set based on the fact that Fuzzy membership function and Fuzzy membership value for the complement of a Fuzzy set are two different things. Accordingly, we have already established a new Similarity measure with the help of extended definition of complementation using reference function. In this paper, an effort has been put forward to show the validity of the results obtained from our proposed measure with the help of traditional Hamming distance and Euclidean distance measures.

Keywords: Complement of a Fuzzy set, Fuzzy membership function, Fuzzy membership value, Fuzzy reference function, Fuzzy set, Similarity measure.

1. INTRODUCTION

Zadeh [1] introduced Fuzzy set in 1965. Since Zadeh initiated Fuzzy sets, many approaches and theories treating imprecision and uncertainty have been proposed. Different researchers have proposed different Similarity measures for Fuzzy sets, all based on Zadehian concept. Zadeh defined Fuzzy set in the manner where it has been believed that the classical set theoretic axioms of exclusion and contradiction are not satisfied for Fuzzy sets. Regarding this, Baruah [2,3] proposed that two functions, namely Fuzzy membership function and Fuzzy reference function are necessary to represent a Fuzzy set. As a result, Baruah [2, 3] reintroduced the notion of complement of a Fuzzy set in a way that the set theoretic axioms of exclusion and contradiction can be seen valid for Fuzzy sets also. Neog and Sut [4] have generalized the concept of complement of a Fuzzy set introduced by Baruah[2,3] when the Fuzzy reference function is not zero and defined arbitrary Fuzzy union and intersection extending the definition of Fuzzy sets given by Baruah [2, 3]. As a consequence of which, the previously existing similarity measures of Fuzzy set, which are based on the traditional Zadehian definition of complementation, have appeared illogical. Accordingly, we have proposed a new Similarity measure[10] for Fuzzy sets using the extended definition of complementation[2,3,4] based on reference function so that it becomes free from any further controversy. In this article, our purpose is to prove that the results obtained from the application of our proposed measure are absolutely valid with respect to traditional Hamming distance and Euclidean distance measures.

The overall organization of this paper is as follows. In section 2 we discuss the new Similarity measure for Fuzzy sets based on extended definition of complementation. In section 3 we apply the new Similarity measure to calculate similarity measures of some collected data. In section 4 we verify the results obtained in section 3 with the results obtained by using Hamming distance and Euclidean distance measures. Finally, some conclusions are given in section 5.

2. THE NEW SIMILARITY MEASURE FOR FUZZY SETS

The new Similarity Measure [10] for Fuzzy sets with the extended definition of complementation is as follows:

Let A and B be two elements belonging to a Fuzzy set (or sets) .Now we can measure the similarity between A and B as below:

$$\text{Sim}(A, B) = \frac{l_{FS}(A,B)}{l_{FS}(A,B^C)} = \frac{a}{b} \qquad (1)$$

where a is distance from $A(\mu_m,\mu_r,\mu_v)$ to $B(\mu_m,\mu_r,\mu_v)$ and b is a distance from $A(\mu_m,\mu_r,\mu_v)$ to $B^C(\mu_m,\mu_r,\mu_v)$ where μ_m, μ_r, μ_v are membership function, reference function and membership value respectively.

For this similarity measure, we have,

$$0 \leq \text{Sim}(A, B) \leq \infty$$

Similarly we can calculate the Similarity between two Fuzzy sets:
Let A and B be two Fuzzy sets defined on the same set of universe of discourse. Now we can measure the similarity between A and B by assessing similarity of the corresponding elements belonging to A and B, as defined in the eqn (1).

Now using Baruah's definition of Fuzzy set, for the Similarity measure of A and B, we can obtain the following 4 possibilities,

A and B may be two exactly similar sets.
or A and B^C may be two exactly similar sets.
or A may be more similar to B than to B^C.
or A may be more similar to B^C than to B.
But A can never be similar to B and B^C together i.e. $A=B=B^C$ is never possible according to the new definition of complementation of Fuzzy set [2, 3].

Therefore from the above analysis, for the Similarity measure of A and B, we can conclude four possible cases as follows:

Case 1: Sim(A,B)=0 when A=B i.e. AB=0.
Case 2: Sim(A,B)=∞ when A=B^C i.e. AB^C=0.
Case 3: Sim(A,B) >1 when AB>AB^C.
Case 4: Sim(A,B) <1 when AB<AB^C.

Hence to measure the similarity between the two Fuzzy sets A and B, one should be interested in the values 0 $\leq$ Sim(A,B) $<$1.

Let us explain the above idea for the new Similarity measure into details:

Let A and B be two Fuzzy sets defined on the same set of universe of discourse U= {e_1, e_2, e_3, e_4, e_5}. Now we can calculate the similarity measure for A and B assessing the similarity measure for the every corresponding elements of A and B i.e. for the every element e_1,e_2,e_3,e_4,e_5 of the set of universe of discourse U, considered for A and B. This means similarity measure for A and B has to be calculated with respect to every e_1, e_2, e_3, e_4, e_5 $\in$ U.

Now, based on the new definition of Fuzzy set, the similarity measure for the Fuzzy set A(e_k, k=1,2,3,4,5) and the Fuzzy set B(e_k , k=1,2,3,4,5) can be obtained under the 3 possible cases in the following manner:

We can visualize the Fuzzy set A (e_k) and the Fuzzy set B (e_k) in the number line in Figure 1 and Figure 2 respectively.

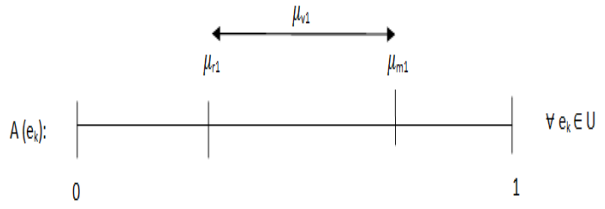


Figure 1. Representation of Fuzzy set A(e_k) in number line

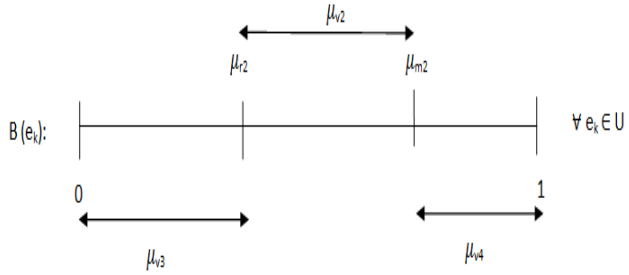


Figure 2. Representation of Fuzzy set B(e_k) in number line

Where μ_{r1}, μ_{m1}, μ_{v1} ; μ_{r2}, μ_{m2}, μ_{v2} ; 0, μ_{r2}, μ_{v3} ; μ_{m2},1,μ_{v4} are reference function, membership function and membership value of the Fuzzy set A , the Fuzzy set B and the two complement sets of B respectively for every e_k $\in$ U.

Now the 3 possible cases are:

Case 1: when $\mu_{r2} \neq 0$, $\mu_{m2} \neq 1$.

Case 1 can be visualized in Figure 1 and Figure 2 and Similarity Measure can be defined as,

$$\frac{AB}{AB^C} =$$

$$Sim(A,B) = \frac{1}{5}\sum_{k=1}^{5}\left(\frac{|A(S_k(\mu_{r1}))-B(S_k(\mu_{r2}))|+|A(S_k(\mu_{m1}))-B(S_k(\mu_{m2}))|+|A(S_k(\mu_{v1}))-B(S_k(\mu_{v2}))|}{|A(S_k(\mu_{r1}))-0|+|A(S_k(\mu_{r1}))-B(S_k(\mu_{m2}))|+|A(S_k(\mu_{m1}))-B(S_k(\mu_{r2}))|+|A(S_k(\mu_{m1}))-1|+|A(S_k(\mu_{v1}))-B(S_k(\mu_{v3}))|+|A(S_k(\mu_{v1}))-B(S_k(\mu_{v4}))|}\right)$$

Case 2: when $\mu_{r2}=0$, $\mu_{m2} \neq 1$.

Case 2 can be visualized in Figure 3 and Figure 4.

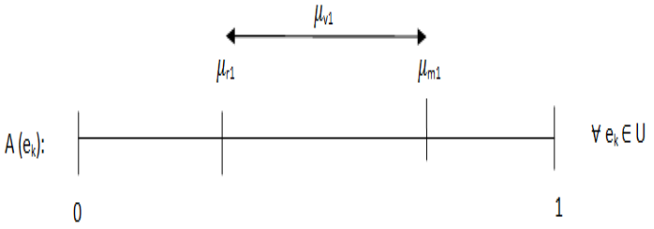


Figure 3. Representation of Fuzzy set A(e_k) in number line

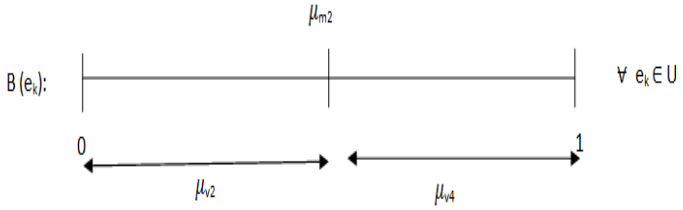


Figure 4. Representation of Fuzzy set B(e_k) in number line

and Similarity Measure can be defined as,

$$\frac{AB}{AB^C} =$$

$$Sim(A,B) = \frac{1}{5}\sum_{k=1}^{5}\left(\frac{|A(S_k(\mu_{r1}))-0|+|A(S_k(\mu_{m1}))-B(S_k(\mu_{m2}))|+|A(S_k(\mu_{v1}))-B(S_k(\mu_{v2}))|}{|A(S_k(\mu_{r1}))-B(S_k(\mu_{m2}))|+|A(S_k(\mu_{m1}))-1|+|A(S_k(\mu_{v1}))-B(S_k(\mu_{v4}))|}\right)$$

Case 3: when $\mu_{r2} \neq 0$, μ_{m2}=1.

Case 3 can be visualized in Figure 5 and Figure 6.

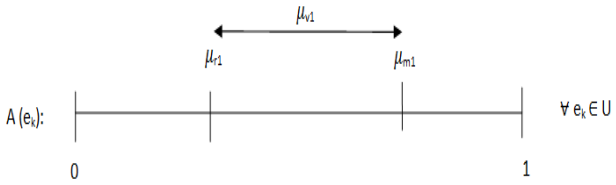


Figure 5. Representation of Fuzzy set A(e_k) in number line

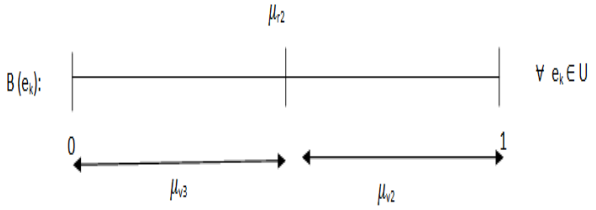


Figure 6. Representation of Fuzzy set B(e_k) in number line

and Similarity Measure can be defined as,

$$\frac{AB}{AB^C} =$$

$$Sim(A,B) = \frac{1}{5}\sum_{k=1}^{5}\left(\frac{|A(S_k(\mu_{r1})) - B(S_k(\mu_{r2}))| + |A(S_k(\mu_{m1})) - 1| + |A(S_k(\mu_{v1})) - B(S_k(\mu_{v2}))|}{|A(S_k(\mu_{r1})) - 0| + |A(S_k(\mu_{m1})) - B(S_k(\mu_{r2}))| + |A(S_k(\mu_{v1})) - B(S_k(\mu_{v3}))|}\right)$$

3. APPLICATION OF THE NEW SIMILARITY MEASURE

We apply our proposed measure on some collected data[5].

Taking larger value (membership or non-membership function value) [5] as membership function value, μ_v and the smaller value[5] as reference function value, μ_r (since, always $0 \leq \mu_r(x) \leq \mu_v(x) \leq 1$), we can represent the collected data[5] as dataset1={A_1,A_2, A_3, A_4, A_5} and dataset2={B_1,B_2,B_3,B_4} defined on the same set of universe of discourse U={a, b, c, d, e} as given in Table 1 and Table 2.

Table 1: dataset 1 = {A_1, A_2, A_3, A_4, A_5}.

	A_1	A_2	A_3	A_4	A_5
a	(0.7,0.1)	(0.8,0.1)	(0.9,0.1)	(0.7,0.2)	(0.8,0.1)
b	(0.4,0.3)	(0.7,0.0)	(0.6,0.2)	(0.7,0.2)	(0.8,0.2)
c	(0.7,0.1)	(0.9,0.0)	(0.7,0.2)	(0.8,0.0)	(0.8,0.2)
d	(0.5,0.3)	(0.6,0.2)	(0.6,0.1)	(0.4,0.2)	(0.8,0.0)
e	(0.4,0.0)	(0.7,0.0)	(0.3,0.3)	(0.7,0.1)	(0.8,0.1)

Table 2: dataset 2 = {B_1, B_2, B_3, B_4}.

	B_1	B_2	B_3	B_4
a	(0.6,0.1)	(0.8,0.1)	(0.5,0.0)	(0.4,0.3)
b	(0.6,0.1)	(0.7,0.1)	(0.7,0.2)	(0.7,0.2)
c	(0.8,0.2)	(0.6,0.1)	(0.6,0.0)	(0.4,0.3)
d	(0.6,0.1)	(0.4,0.4)	(0.8,0.1)	(0.5,0.4)
e	(0.8,0.1)	(0.8,0.0)	(0.8,0.1)	(0.6,0.1)

Each element (a, b, c, d or e) ∈ U in Table1 and Table2 is described by: a reference function and a membership function value.

Now to calculate a similar set from the dataset 1 for a particular set in dataset 2, we proceed in the following way:

Step 1: At first we calculate the similarity measure $\frac{B_jA_i}{B_jA_i^C}$ for each set B_j ∈ dataset 2,(where j=1,2,3,4) with every set A_i ∈ dataset 1 ,where (i=1,2,3,4,5) separately ,assessing the similarity measure for the every corresponding elements of the two sets i.e. a,b,c,d,e ∈ U, the set of universe of discourse considered for the two datasets.

Step 2: Then we find out the smallest value from the obtained similarity measures between a set B_j and every set A_i, we considered in Step 1.From that value we can decide which A_i ∈ dataset 1 is similar to a particular set B_j ∈ dataset 2 .

Now we calculate the similarity measure values between the dataset 1 and the dataset 2 and represent the calculated values in table 3.

Table 3: Similarity measure values between dataset 1 and dataset 2.

	A_1	A_2	A_3	A_4	A_5
B_1	0.49	0.23	0.28	0.19	0.15
B_2	0.41	0.24	0.37	0.20	0.33
B_3	0.54	0.28	0.35	0.21	0.12
B_4	0.45	0.61	0.49	0.38	0.57

Hence from table 3 we can conclude that,

Set B_1 is similar to set A_5, set B_2 is similar to set A_4, set B_3 is similar to set A_5 and set B_4 is similar to set A_4.

4. VERIFICATION

We can verify the results obtained in section 3 by using normalized Hamming Distance and Euclidean Distance measures.

For that purpose, we apply Hamming Distance measure and Euclidean Distance measure on the same collected data[5] , used in section 3.

4.1 Hamming Distance measure:

To measure similarity between dataset 1 and dataset 2 using Hamming Distance measure, we proceed in the following way:

Step 1: At first we calculate the Hamming Distance measure for each set B_j ∈ dataset 2, (where j=1,2,3,4) with every set A_i ∈ dataset1 ,(where i=1,2,3,4,5) separately ,assessing the

distance measure for the every corresponding elements of the two sets i.e. a,b,c,d,e ∈ U, the set of universe of discourse considered for the two datasets.

Step 2: Then we find out the smallest value from the obtained measures between a set B_j and every set A_i , we considered in Step 1. From that value we can decide which $A_i \in$ dataset 1 is similar to a particular set $B_j \in$ dataset 2 .

The normalized Hamming distance of the set A_i from the set B_j, for all a, b, c, d, e ∈ U is

$$HD(A_i,B_j)=\frac{1}{10}\sum_{k=1}^{5}\left(\left|A_i\left(s_k(\mu_{r1})\right)-B_j\left(s_k(\mu_{r2})\right)\right|+\left|A_i\left(s_k(\mu_{m1})\right)-B_j\left(s_k(\mu_{m2})\right)\right|+\left|A_i\left(s_k(\mu_{v1})\right)-B_j\left(s_k(\mu_{v2})\right)\right|\right)$$

Table 4: Hamming Distance measures between dataset 1 and dataset 2.

	A_1	A_2	A_3	A_4	A_5
B_1	0.26	0.18	0.24	0.16	0.14
B_2	0.26	0.20	0.34	0.20	0.24
B_3	0.32	0.24	0.32	0.20	0.14
B_4	0.34	0.42	0.42	0.28	0.42

From table 4 we conclude that

Set B_1 is similar to set A_5, set B_2 is similar to set A_2 and A_4, set B_3 is similar to set A_5 and set B_4 is similar to set A_4.

4.2 Euclidean Distance measure:

To measure similarity between dataset 1 and dataset 2, we proceed in the following way:

Step 1: At first we calculate the Euclidean Distance measure for each set $B_j \in$ dataset 2, (where j=1,2,3,4) with every set $A_i \in$ dataset1 ,(where i=1,2,3,4,5) separately ,assessing the distance measure for the every corresponding elements of the two sets i.e. a,b,c,d,e ∈ U, the set of universe of discourse considered for the two datasets.

Step 2: Then we find out the smallest value from the obtained measures between a set B_j and every set A_i , we considered in Step 1. From that value we can decide which $A_i \in$ dataset 1 is similar to a particular set $B_j \in$ dataset 2 .

The Euclidean distance of the set A_i from the set B_j, for all a, b, c, d, e ∈ U is

$$ED(A_i,B_j)=\frac{1}{5}\sum_{k=1}^{5}\sqrt{\left(\left|A_i\left(s_k(\mu_{r1})\right)-B_j\left(s_k(\mu_{r2})\right)\right|^2+\left|A_i\left(s_k(\mu_{m1})\right)-B_j\left(s_k(\mu_{m2})\right)\right|^2+\left|A_i\left(s_k(\mu_{v1})\right)-B_j\left(s_k(\mu_{v2})\right)\right|^2\right)}$$

Table 5: Euclidean Distance measures between dataset 1 and dataset 2.

	A_1	A_2	A_3	A_4	A_5
B_1	0.32	0.23	0.31	0.21	0.17
B_2	0.33	0.25	0.42	0.25	0.29
B_3	0.39	0.31	0.40	0.26	0.18
B_4	0.42	0.52	0.52	0.34	0.53

From table 5 we conclude that

Set B_1 is similar to set A_5, set B_2 is similar to set A_2 and A_4, set B_3 is similar to set A_5 and set B_4 is similar to set A_4.

Therefore from table 4 and table 5, we have found that the set B_2 is similar to both the set A_2 and the set A_4. This is due to the limitation that both Hamming distance and Euclidean distance measures can calculate upto two decimal places only. Hence even if there is a fractional difference between two values after two decimal places, these two measures can not represent it. It has been observed that the results obtained from Hamming distance and Euclidean distance are exactly similar. Thus, we have seen that though there is no error in the results, the results obtained from these two traditional measures , sometimes, may not be clear which is not a problem at all in case of our proposed measure. From table 3, we can clearly find that set B_2 is actually similar to set A_4. Also, it has been observed that the results obtained from table 3, table 4 and table 5 are absolutely similar with just a little exception in the result for the set B_2 in case of the two traditional distance measures which determines that our proposed measure gives clearer result than the traditional measures. Therefore, it can be concluded that the results obtained by the application of the new Similarity measure are clear and valid with respect to the traditional Hamming distance and Euclidean distance measures and hence legally acceptable.

5. CONCLUSION

In this work, we have first gone through the new Similarity measure for Fuzzy sets based on the extended definition of complementation based on reference function. Again, the Similarity measure has been applied to evaluate some collected dataset. Also, the results obtained from the application are compared with the results found by Hamming distance and Euclidean distance measures. Finally it has been proved that the results obtained from the application of the new Similarity measure for Fuzzy sets are much clear and absolutely valid with respect to the traditional distance measures.

6. ACKNOWLEDGEMENT

The author would like to thank Hemanta K. Baruah, Professor, Department of Statistics, Gauhati University,for his valuable suggestions and guidance, in preparing this article.

7. REFERENCES

[1] L.A.Zadeh, " Fuzzy Sets" , Information and Control, 8, pp. 338-353, 1965.

[2] Hemanta K. Baruah, "Towards Forming A Field Of Fuzzy Sets" International Journal of Energy, Information and Communications, Vol.2, Issue 1, pp. 16-20, February 2011.

[3] Hemanta K. Baruah, "The Theory of Fuzzy Sets: Beliefs and Realities", International Journal of Energy, Information and Communications, Vol. 2, Issue 2,pp. 1-22, May,2011.

[4]Tridiv Jyoti Neog , Dusmanta Kumar Sut, "Theory of Fuzzy Soft Sets from a New Perspective", International Journal of Latest Trends in Computing, Vol. 2, No 3, September,2011.

[5] Eulalia Szmidt, Janusz Kacprzyk, "Medical Diagnostic Reasoning Using a Similarity Measure for Intuitionistic Fuzzy Sets" Eighth Int. Conf. on IFSs, Varna, 20-21 June 2004, NIFS Vol. 10 (2004), 4, 61-69.

[6]Hong-mei Ju , Feng-Ying Wang , "A Similarity Measure for Interval –valued Fuzzy Sets and Its Application in Supporting Medical Diagnostic Reasoning", The Tenth International Symposium on Operation Research and Its Applications(ISORA 2011), Dunhuang, China, August 28-31, 2011.

[7] Szmidt E., Kacprzyk J. (2001), "Entropy for intuitionistic Fuzzy sets". Fuzzy Sets and Systems, vol. 118, No. 3, pp. 467-477.

[8]Hongmei Ju (2008). "Entropy for Interval-valued Fuzzy Sets",Fuzzy information and engineering, Volume 1,358-366.

[9] Pranamika Kakati (2013). "A Study on Similarity Measure for Fuzzy Sets" International Journal of Advanced Research in Computer Science and Softaware Engineering, vol. 3 Issue 8, pp.97-103, August 2013.

[10]Pranamika Kakati (2013) "A New Similarity Measure for Fuzzy Sets with the Extended Definition of Complementation"International Journal of Soft Computing and Engineering, Volume 3 Issue 4, pp.203-207, September, 2013.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 606 - 608, 2013, ISSN: 2319–8656

Increasing efficiency in Grid by using glow worm Algorithm

Sedigheh Navaezadeh
Department of Computer Science and Research Branch
Islamic Azad University
Khouzestan, Iran

Mehdi Sadeghzadeh
Member of Science Board of Computer group in Azad
Islamic university of Mahshahr.

Ali Harounabadi
Member of Science Board of Computer Group in Azad
Islamic university of Tehran center.

Abstract: Nowadays, human beings encounter with hug data. Due to development of computer and detector technologies, few terabyte data has been produced. The analysis of these information require the sources whose financial costs cannot be afforded by any institution. In order to meet this demand, Grid computing has been considered as one of the most important fields of research. One of the important issues in grid environment is reliability in Grid service which must be studied. The study of Grid service helps systems' designers to create better systems with more reliability. Also, it helps the users to use Grid environment logically. The main purpose of Grid system is preparing services with high reliability and the least cost for many users as well as supporting collaborative jobs. On one hand, resources have important role in increasing and decreasing Grid application. If resources are limited, defilewii be the result; therefore, will decrease Gird efficiency will decrease. In this paper, swarm intelligence Algorithm has been applied for Grid load balance. Since Gird environment is dynamic, we use glowworm algorithm in sites in order to balance the load. The algorithm simulation results presented by MATLAB software showed that this algorithm can increase efficiency as well as Grid environment reliability.

Keywords: swarm intelligence algorithm, glowworm algorithm, Grid, load balance.

1. INTRODUCTION

Solving many problems in the field of engineering and other sciences requires using powerful computing resources. Computing Grid environments are suitable beds for solving problems which need long and difficult computing. Resources have been geographically distributed in this environment, but they can be logically considered as a unit source. Grid computing systems have been created as an important and modern field, and are different from common distributed systems. This is due to source subscription in large scales and innovative applicable programs with high efficiency. Grid computing system is a kind of distributed systems with a broad range. The main and special problem of Grid concept is resource subscription coordination, dynamic problem solving and multi- distributed programs.

Subscription in which Grid computing is involved is not only file primary exchange but also direct accessibility to computers, software data and other sources needing a broad spectra of problem solving strategies and resource subscription. Therefore, Grid computing reliability has been influenced by the relation existing between computing programs and important sources [1].

Since scheduling problem is related to NP-hard issue, we can use the assessment algorithm to solve it. Among these algorithms, genetic algorithm (GA) is more common, so it has been used in this paper.

GA presents new solutions through mixing the best options and existing solutions randomly. In this algorithm, the function called intersection has been used for general quest among the solutions by means of changing a part of 2 chromosomes randomly. Another important local research operator is mutation in which random change of one gene in a chromosome has been used and applied. Whole process that has been repeated several times is called propagation or reproduction [2]. The advantages of computing Grid are as the following: sources, parallel computing ability, creating source and virtual organization, using source, reliability, management and more source availability [3].

Since artificial intelligence algorithms are efficient in optimization, they have been considered as good options for solving the problems such as load balance in distributed system. The purpose of load balance is decreasing the rate of difference between the heaviest node and the lightest node in terms of load rate.

2. LITERATURE

Load Balance algorithm is divided into different groups. On one hand, these algorithms are divided into two groups; namely, static and dynamic. The difference of load balance in dynamic and static mode is that, in static mode, the decisions related to load balance are made in compiling time, while, in dynamic mode, decisions related to load balance are made in performing time. That is, in static mode, these decisions are made in the time of request for the source, but, in dynamic mode, the behavior of balancer varies according to the changes of parameters and policies [5].

Load balancing is divided into three groups: concentrated, non – concentrated and hierarchy. In concentrated method, all functions are scheduled by a scheduler, and scheduling operation is performed by the applicant source. In hierarchy method, the scheduler is organized in the hierarchy form [6]

In load balance of concentrated method, many studies have been conducted. In order to balance the load, genetic algorithm has been used by [5], and its Simulation result has been compared with Min – Max and Max – Min algorithm. In [2], a new genetic algorithm has been presented by using resource fault occurrence history (RFOH) for certain scheduling in computing Grid.

This strategy saves source fault occurrence history in Gird information server. GA is a general quest technique in which complex probable solutions called chromosomes have been applied and used. GA has created new solutions through mixing the best options among existing solutions randomly. The use of this information has decreased selection chance of the resources having more failure probability.

Simulation results showed that proposed strategy decreased total time of programs performance. One of glowworm algorithm advantages relating to other swarm intelligence algorithms is its constriction simplicity.

Firefly Algorithm is a type of algorithm obtained from nature and collective smart algorithm which presented by yang in

2008. This algorithm is a modern technique based on collective behaviors and inspired from firefly social behaviors in the nature. Swarm intelligence is a type of artificial intelligence based on collective behaviors and neutralized and self-organized foundations. Fireflies generate rhythmic and short beams. Optical patterns of each firefly are different from others. Fireflies use these beams for two reasons: pairs attraction process, and attracting hunt. Moreover, these beams are used as a protective mechanism for fireflies. Due to rhythmic beams and rate of radiation and interval rate between light signals, two genders attract each other. Each particle of the firefly in multidimensional quest space is updated by dynamic absorption and based on the knowledge of firefly and its neighbors.

Firefly optimization algorithm can be stated as follows [7]:

* All fireflies are single- gender, and the factor of pairs attractiveness are considered regardless of their gender.

* Firefly x attracts all fireflies, and is attracted by all fireflies.

* Attractiveness is related to their glow, so in a pair of firefly, a worm with less light is attracted toward a worm with more light. Attractiveness power is related to their beam, and the light intensity decreases when the distance between two fireflies increases. If a firefly is not brighter than others, their movement will be performed randomly.

* Brighter firefly moves randomly (all fireflies can not attract them).

* Firefly brightness is determined by objective function value. Light intensity can be easily determined by target function.

* Firefly particles are randomly distributed in quest space according to above principles. There are two main parts in firefly algorithm: attracting firefly and moving toward the attracted firefly.

2.1 General Form Of Firefly Algorithm

General form of firefly algorithm has been shown in figure (1). As it can be observed in this figure, primary coordination and light intensity rate and the distance between firefly particles are firstly determined in quest area. In Quest procedure of firefly algorithm. Each firefly is individually compared with other fireflies. If the firefly has less light than the compared one, it will move toward the firefly with more light (the problem of finding maximum point), and due to this process, particles are centralized around a particle with more light. In the next generation of algorithm, if there is a particle with more light, then the particles will again move toward the particle with more light. Quest stages must be generated according to maximum number of repetitions. In this paper, in order to optimize the problem of load balance through using glowworm group intelligence algorithm, a solution has been presented. In this method, each node in the network is considered as a glowworm. Each glowworm tries to optimize its own existing load rate, and this work can be done by exchanging the load with other nodes.

3.THE PROPOSED ALGORITHM

The application of glowworm algorithm has been explained in previous sections. In the algorithm presented on the basis of swarm intelligence, all nodes in Grid system are considered as a solution for finding the most optimized mode.

The place of each node has been shown by the existing rate of light. In order to determine the attraction parameter for each node, its node light rate is compared with neighbor nodes. Each node always moves toward the best neighbor. This work can be done by attracting toward the neighbor or emitting more light from neighbor nodes when the rate of a node light is as same as the rate of neighbor nodes. The movement is performed randomly.

Since grid is a dynamic environment, solutions are always changing; therefore,, in this method, there is no need to keep information and previous history like classic glow worm . In addition, no massive particles are required. In this research, parameter α , β and r are considered from 0 to 1.

Particle Algorithm()
sourceLoad
while running
Do if job Queue.size>0
Then
Lighteload %% choose best Neighbor(entekhab avalin behtarin)
SecondLightestLoad %% choose SecondLightest Neighbor(en entekhab dovomin behtarin)
TC (kamtarin hazine tebghe kamtarin faseleh)
threshold(tebghe avalin va dovomin)
while tc>threshold
do
Submit jobs %%(TC)
sourceLoad %%(currentNodeload)
velocity %%(sourceload- lightestLoad)

Figure 1: Running Algorithm by using glow worm.

3.1 Simulation Results

After the proposed algorithm simulation in MATLAB environment, the results are compared with massive particles and genetic. The procedure of creating the network topology for the simulation is as follows:

At first, a minimum covering tree is structured, and then topology is obtained by adding the edges. In figure (2), time of the first job sending to the network and the last job are compared. The diagram showed the application of glowworm algorithm is more optimized than two other algorithms.

In figure (3), the difference between the lightest and heaviest nodes have been indicated in terms of the load (the load in glow worm shows its own light, and due to its similarity with massive particles, we have shown job comparison with this algorithm.). In figure (4), the effects of increasing number of jobs on time have been shown. This time involves sending time of the first job and running the last job in the network.

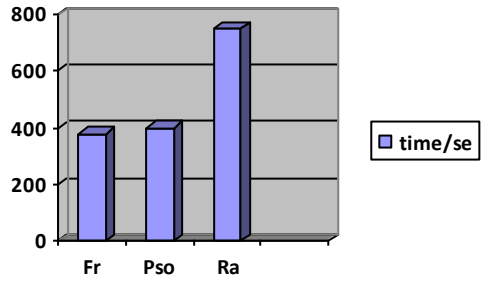


Figure 2: the advantage of running time in different algorithms

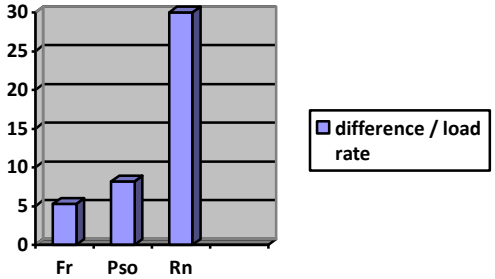


Figure 3: the difference between load rate of the worm with less light and the worm with more light.

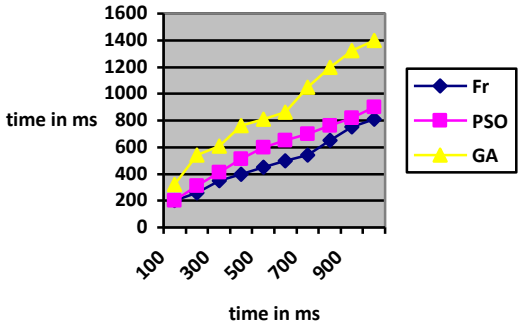


Figure 4: increasing the number of jobs along with increasing the number of jobs time.

4. CONCLUSION

In this paper, glow worm algorithm has been considered in order to balance the load in Grid. The results showed that this algorithm is more efficient than other algorithms when scheduling is considered, and there are different sources. Glow worm algorithm showed better results in terms of the running time. In addition, glowworm algorithm has been considered as the most appropriate methods for optimization problems.

5. REFERENCES

[1] Dai .Y.S, . Xie. M, Poh .K.L. (2010),.," Reliability Analysis of Grid Computing Systems", Department of Industrial and System Engineering, National University of Singapore.

[2] Mohammad Khanli. Leyli, Etminan Far . Maryam, Ghaffari . Ali. 2010," Reliable Job Scheduler using RFOH in Grid Computing", Journal of Emerging Trends in Computing and Information Sciences CIS Journal. All rights reserved.

[3] Cao. J, Spooner. D. P, Jarvis. S. A, Nudd.v.,(2005)._Grid load balancing using intelligent agents. Future_Gener. Comput. System.

[4] Grosan. C, Abraham. A, and Helvik. B. (2007).Multiobjective evolutionary algorithms for scheduling jobs on computational grids.ADIS InternationalConference, Applied Computing, Salamanca,Spain, Nuno Guimares and Pedro Isaias (Eds).

[5] Subrata. R., Zomaya. A. Y, and. Landfeldt. B(2007). Arti_cial_ life techniques for load balancing in computational_grids. J. Comput. Syst. Sci.,

[6] J. K. J and. Eberhart. R. C,(2001). Swarm Intelligence._Morgan Kaufmann Publishers

[7] Hassanzadeh, T. Meibodi, M, 1390. Presenting Improved Glowworm Algorithm for Optimization in Static Environment. The Fifth Conference of Datamining in Iran. Amirkabir University, Iran:Tehran..

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 609 - 613, 2013, ISSN: 2319–8656

Gene Selection for Cancer Classification using Microarrays

T.Shanmugavadivu,
Karpagam University,
Coimbatore-21

T.Ravichandran
Hindusthan Institute of Technology,
Coimbatore-32.

Abstract Microarrays allow biologists to better understand the interactions between diverse pathologic states at the gene level. However, the amount of data generated by these tools becomes problematic. New techniques are then needed in order to extract valuable information about gene activity in sensitive processes like tumor cells proliferation and metastasis activity. Recent tools that analyze microarray expression data have exploited correlation-based approach such as clustering analysis. Here wed scribe a novel GA/ANN distributed approach for assessing the importance of genes for sample classification based on expression data. Several different approaches have been exploited and a comparison has been given. The developed system was employed in the classification of ER+/- metastasis recurrence of breast cancer tumors and results were validated using a real life database. Further validation has been carried out using Gene Ontology based tools. Results proved the valuable potentialities and robustness of similar systems.

Key Words : Diagnosis, diagnostic tests, drug discovery, Support Vector Machines.

1. INTRODUCTION

Introduced for the first time in 1989, microarrays have gained in this time a great fame thanks to their ability to give biologists a quite detailed snapshot of cellular and genomic activity in particular states of the examined organism. Recent advances in microarray technology have allowed studying the expression patterns of thousands of genes in parallel. The principles these devices are based on are really few and simple. Microarrays use hybridization-based methodology that allows mRNA molecules to bind to their complementary parts (genes). Several probes for each gene are placed on a coated quartz surface (1.28 cm x 1.28/ cm); mRNA segments hybridize with probes according to A-T C-G base pairing principle and this allows the monitoring of the expression levels of thousands of genes simultaneously. This enables the measurement of the levels of mRNA molecules inside a cell and, consequently, the proteins being produced. Hence, the role of the genes in a cell at a given moment can be better understood by analyzing their expression levels. In this context, the comparison between gene expression patterns through the measurement of the levels of mRNA in healthy versus unhealthy cells can supply important information about pathological states, as well as information that can lead to earlier diagnosis and more efficient treatment.

Many genes are strongly regulated and only transcribed at certain times, in certain environmental conditions, and in certain cell types. Microarrays simultaneously measure the mRNA expression level of thousands of genes in a cell mixture. By comparing the expression profiles of different tissue types we might find the genes that best explain a perturbation or might even help clarify how cancer is developing. Given a series of microarray experiments for a specific tissue under different conditions. To find the genes most likely differentially expressed under these Conditions. In other words, To find the genes that best explain the effects of these conditions. This task is also called feature selection, a commonly addressed problem in machine learning, where one has class-labeled data and wants to figure out which features best discriminate among the classes. If the genes are the features describing the cell, the problem is to select the features that have the biggest impact on describing the results and to drop the features with little or no effect. These features can then be used to classify unknown data. Noisy or irrelevant attributes make the classification task more complicated, as they can contain random

correlation. Therefore we want to filter out these features.

2. GENE IDENTIFICATION:

The dataset contains gene expression information extracted from DNA microarrays. This microarray dataset is used to distinguish tumor and normal tissues. There are 62 tissue samples, of which 22 are normal and 40 are cancer tissues, each having 2000 genes with highest minimal intensity across the 62 issues. The data set was divided into a training set with 32 samples and a test set with 30 samples total number of 5157 subsets of genes that correctly classify all training samples are obtained using our bootstrapped GA/SVM algorithms. Each subset consists of five genes. Genes are then ordered based on the number of occurrences with which genes are selected. Figure 1shows the number of occurrences for each gene.

3. GENE CLASSIFICATION:

Classification problems where the input is a vector that call a "pattern" of n components and call a "features". F the n dimensional feature space. In the case of the problem at hand, the features are gene expression coefficients and patterns correspond to patients. The limit ourselves to two-class classification problems. To identify the two classes with the symbols (+) and (-). A training set of a number of patterns $\{x_1, x_2, \dots x_k, \dots x_l\}$ with known class labels $\{y_1, y_2, \dots y_k, \dots y_l\}$, y_kî{-1,+1}, is given. The training patterns are used to build a decision function (or discriminate function) D(x), that is a scalar function of an input pattern x. New patterns are classified according to the sign of the decision function:

D(x) > 0 =>class (+)
D(x) < 0 => class (-)
D(x) = 0, decision boundary.

Decision functions that are simple weighted sums of the training patterns plus a
bias are called linear discriminate functions In our notations:

D(x) = w.x+b, (1)
where w is the weight vector and b is a bias value.
A data set is said to be "linearly separable" if a linear discriminate function can
separate it without error.

3. SPACE DIMENSIONALITY REDUCTION AND FEATURE SELECTION:

A known problem in classification specifically, and machine learning in general, is to find ways to reduce the dimensionality n of the feature space F to overcome the risk of "over fitting". Data over fitting arises when the number n of features is large (in our case thousands of genes) and the number l of training patterns is comparatively small (in our case a few dozen patients). In such a situation, one can easily find a decision function that separates the training data (even a linear decision function) but will perform poorly on test data. Training techniques that use regularization avoid over fitting of the data to some extent without requiring space dimensionality reduction. Such is the case, for instance, of Support Vector Machines (SVMs) benefit from space dimensionality reduction.

4. SUPPORT VECTOR MACHINE (SVM) ALGORITHM:

The SVM learning algorithm is fairly simple. Our implementation follows the formulation .This approach differs slightly from that the geometric interpretation remains the same.

$$\mathbf{L(x)}=\sum_{i=1}^{n} y_i \propto_i k(\mathrm{x}, \mathrm{x}_i)$$

The goal is to learn a set of weights that maximize the following objective function:

$$J(\propto) = \sum_{i=1}^{n} \propto_i (2 - y_i L(x_i))$$

$$= f\left(\frac{1 - y_i L(x_i) + \propto_i k(x_{i,} x_i)}{k(x_{i,} x_i)}\right)$$

This maximum can be obtained by iteratively updating the weights using the following update rule:

$$\propto_i \leftarrow f\left(\frac{1 - y_i L(x_i) + \propto_i k(x_{i,} x_i)}{k(x_{i,} x_i)}\right)$$

where $f(x)=x$ for $x > 0$ and $f(x)=0$ for $x <= 0$. difference arises because we implement the soft margin by

modifying the diagonal of the kernel matrix, rather than by truncating the weights.

5. FEATURE RANKING WITH SUPPORT VECTOR MACHINES:

To test the idea of using the weights of a classifier to produce a feature ranking, we used a state-of-the-art classification technique: SVMs have recently been intensively . They are presently one of the best-known classification techniques with computational advantages over their contenders SVMs. The handle non-linear decision boundaries of arbitrary complexity, we limit ourselves, in this paper, to linear SVMs because of the nature of the data sets under investigation. Linear SVMs are particular linear discriminate classifiers An extension of the algorithm to the non-linear If the training data set is linearly separable, a linear SVM is a maximum margin classifier. The decision boundary (a straight line in the case of a two-dimensional separation) is positioned to leave the largest possible margin on either side.

6. COLON CANCER DIAGNOSIS:

Gene expression information was extracted from DNA micro-array data . The 62 tissues include 22 normal and 40 colon cancer tissues. The matrix contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Some genes are non-human genes provides an analysis of the data based on top down hierarchical clustering, a method of unsupervised learning. They show that most normal samples cluster together and most cancer samples cluster together. They explain that “outlier” samples that are classified in the wrong cluster differ in cell composition from typical samples. They compute a so-called “muscle index” that measures the average gene expression of a number of smooth muscle genes.

7. RESULTS :

Our experiments show the benefits of classifying genes using support vector machines trained on DNA microarray expression data. We begin with a comparison of SVMs versus four non-SVM methods and show that SVMs provide superior performance. We then examine more closely the performance of several different SVMs and demonstrate the superiority of the radial basis function SVM. Finally, we examine in detail some of the apparent errors made by the radial basis function SVM and show that many of the apparent errors are in fact biologically reasonable classifications.

SVM performance using various kernels. SVMs were trained using four different kernel functions on five different random three-fold splits of the data, training on two-thirds and testing on the remaining third. The first column contains the class . The second column contains the kernel function.The next five columns contain the threshold-optimized cost (i.e., the number of false positives plus twice the number of false negatives) for each of the five random three-fold splits. The final column is the total cost across all five splits.

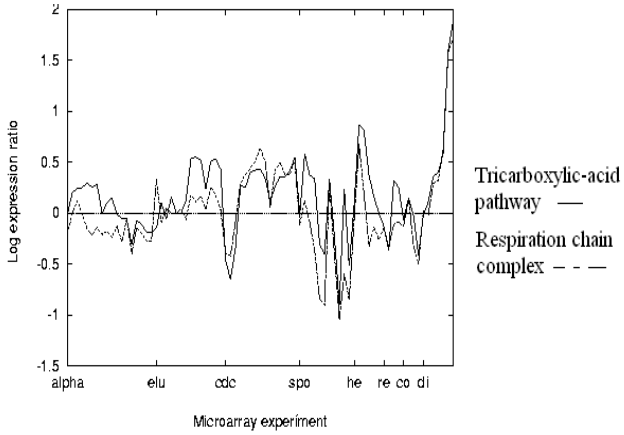
Class	Kernel	Cost for each split					Total
Tricarboxylic acid	Radial	18	21	15	22	21	97
	Dot-product-1	15	22	18	23	22	100
	Dot-product-2	16	22	17	22	22	99
	Dot-product-3	16	22	17	23	22	100
Respiration	Radial	16	18	23	20	16	93
	Dot-product-1	24	24	29	27	23	127
	Dot-product-2	19	19	26	24	23	111
	Dot-product-3	19	19	26	22	21	107
Ribosome	Radial	8	12	15	11	13	59
	Dot-product-1	13	18	14	16	16	77

	Dot-product -2	11	16	14	16	15	72
	Dot-product -3	9	15	11	15	15	65
Proteasome	Radial	14	10	9	11	11	55
	Dot-product -1	16	12	12	17	19	76
	Dot-product -2	16	13	15	17	17	78
	Dot-product -3	16	13	16	16	17	79
Histone	Radial	4	4	4	4	4	20
	Dot-product -1	4	4	4	4	4	20
	Dot-product -2	4	4	4	4	4	20
	Dot-product -3	4	4	4	4	4	20

To demonstrating the superior performance of SVMs relative to non-SVM methods, the radial basis SVM performs better than SVMs that use a scaled dot product kernel. In order to verify this difference in performance, we repeated the three-fold cross-validation experiment four more times, using four different random splits of the data. The total cost in all five experiments is reported in the final column of the table. The radial basis SVM performs better than the scaled dot product SVMs for all classes except the histones, for which all four methods perform identically.

The number of genes that a radial basis SVM misclassifies only once in the five experiments. The right-most column lists the number of genes that are consistently misclassified in all five experiments. These latter genes are of much more interest, since their misclassification cannot be attributed to an unlucky split of the data

Each series represents the average log expression ratio for all genes in the given family plotted as a function of DNA microarray experiment.



These are genes for which the radial basis support vector machine consistently disagrees with the classification. Many of these disagreements reflect the different perspective provided by the expression data concerning the relationships between genes. The microarray expression data represents the genetic response of the cell to various environmental perturbations, and the SVM classifies genes based on how similar their expression pattern is to genes of known function. The definitions of functional classes have been arrived at through biochemical experiments that classify gene products by what they do, not how they are regulated. These different perspectives sometimes lead to different functional classifications. The genes that are regulated at the translational level or protein level, rather than at the trancriptional level measured by the microarray experiments, cannot be correctly classified by expression data alone. Third, genes for which the microarray data is corrupt cannot be correctly classified. Disagreements represent the cases

where the different perspectives of the SVM lead to different functional classifications and illustrate the new information that expression data brings to biology.

8. CONCLUSIONS:

SVMs lend themselves particularly well to the analysis of broad patterns of gene expression from DNA micro-array data. They can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients). They integrate pattern selection and feature selection in a single consistent framework.

The top ranked genes found by SVM all have a plausible relation to cancer. In contrast, other methods select genes that are correlated with the separation at hand but not relevant to cancer diagnosis. This simple method allows us to find nested subsets of genes that lend themselves well to a model selection technique that finds an optimum number of genes. Our explorations indicate is much more robust to data overfitting than other methods, including combinatorial search.

Further work includes experimenting with the extension of the method to nonlinear classifiers, to regression, to density estimation, to clustering, and to other kernel methods. We envision that linear classifiers are going to continue to play an important role in the analysis of DNA micro-array because of the large ratio number of features over number of training patterns. generally, the simultaneous choice of the learning machine and the feature subset should be addressed, an even more complex and challenging model selection problem.

9. REFERENCES:

[1] (Aerts, 1996) Chitotriosidase - New Biochemical Marker. Hans Aerts. Gauchers News, March, 1996.

[2] (Alizadeh, 2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Ash A. Alizadeh *et al*, Nature, Vol. 403, Issue 3, February, 2000.

[3] (Alon, 1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. Alon *et al*, PNAS vol. 96 pp. 6745-6750, June 1999, Cell Biology. The data is available on-line at http://www.molbio.princeton.edu/colondata.

[4] (Aronson, 1999) Remodeling the Mammary Gland at the Termination of Breast Feeding: Role of a New Regulator Protein BRP39, The Beat, University of South Alabama College of Medecine, July, 1999.

[5] (Ben Hur, 2000) A support vector method for hierarchical clustering. A. Ben Hur, D. Horn, H. Siegelman, and V. Vapnik. Submitted to NIPS 2000. (Boser, 1992) An training algorithm for optimal margin classifiers. B. Boser, I. Guyon, and V. Vapnik. In Fifth Annual Workshop on Computational Learning Theory, pages 144--152, Pittsburgh, ACM. 1992.

[6] (Perou, 1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Charles M. Perou et al Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 9212–9217, August 1999, Genetics.

[7] (Schölkopf, 1998) Non-linear component analysis as a kernel eigenvalue problem. B. Schölkopf, A. Smola, K.-R. Muller. Neural computation, vol. 10, pp. 1299-1319, 1998.

[8] (Smola, 2000) Sparce greedy matrix approximation for machine learning. A. Smola and B. Schölkopf. Proceedings of the 17th international conference onmachine learning, pp 911-918. June, 2000.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 614 - 617, 2013, ISSN: 2319–8656

Detect Breast Cancer using Fuzzy C means Techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets

Tintu P B
CMS College of Science and Commerce
Coimbatore, Tamil Nadu, India

Paulin. R
CMS College of Science and Commerce
Coimbatore, Tamil Nadu, India

Abstract- Medical data mining is very much valuable to medical experts. The main task of data mining is diagnosing the patient's disease Classification. Breast cancer is a severe and life threatening disease very commonly found in woman. An unusual growth of cells in breast is the main source of breast cancer those cells can be of two types malignant (Cancerous) and benign (Non-Cancerous) these types must be diagnosed taking proper meditation and for proper treatment. Modern medical diagnosis scheme is totally based on data taken through clinical and/or other test; most of the decision related to classification of a disease is a very crucial and challenging job. In this research work, using intelligent techniques of data mining is Fuzzy C Means; we have focused on breast cancer diagnosis by fuzzy systems. Fuzzy rules are desirable because of their interpretability by human experts. It has been applied to classify data related to breast cancer from UCI repository site. Experimental works were done using MATLAB in order to reduce dimensionality of breast cancer data set a ranking based feature selection technique. Results on breast cancer diagnosis data set from UCI machine learning repository show that this approach would be capable of classifying cancer cases with high accuracy rate in addition to adequate interpretability of extracted rules.

Keyword— Classification, Clustering, Fuzzy C Means, Breast Cancer, Wisconsin Prognostic Breast Cancer (WPBC).

1. INTRODUCTION

Proper diagnosis of any human disease precisely and powerfully is a challenging task for the people involved in health care organization and offers a strong base for further treatment and medication. Breast cancer comes fourth in cancer diagnosis in women between 20 to 29 years. Breast cancer is most common type of cancer in women, with more than one million instances and nearly half million of deaths occurring worldwide annually [1]. In 2010, there were reported approximately 207090 newly diagnosed cases and 30840 deaths in the United States, and total of 1,638,910 new cancer cases is projected to occur in 2012 [2]. A breast cancer victim's chances for long-term survival are improved by early detection of the disease, and early detection is in turn enhanced by an accurate diagnosis. In addition, the National Cancer Institute of U.S. estimates that 16.4 percent of women born today and live with a breast cancer diagnosis [3]. For the diagnosis of the breast cancer cases as well as for the prognosis of the disease many methods have been discussed [4], [5], [6], [7], [8], [9], [10] and [11]. All machine leaning techniques to provide the same levels of exactness, without the negative sides of surgical biopsy.

The biggest problem in medical science includes the diagnosis of disease since the reason of breast cancer is unknown, although scientists know some of the risk factors like ageing, genetic risk factors, menstrual periods, family history, not having children, alcohol, overweight, obesity, etc. [12]. Symptoms of cancer include a lump in the breast or underarm that persists after menstrual cycle, swelling in the armpit, pain or tenderness in the breast, any change in the size, texture, contour, or temperature of the breast, a marble-like area under the skin. Many cancer diseases take place within the pale of the same family and the immediate relatives (siblings, parents, and children) of patients with cancers often have an increased risk of cancer.

This paper deals with the breast cancer diagnosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) data sets, which are publicly available by anonymous ftp [13]. These data sets involve measurements taken permitting the Fine Needle Aspirate (FNA) test. In case that a patient is diagnosed with breast cancer, the malignant mass must be excised. After this or a different post-operative procedure, a prediction of the expected course of the disease must be determined. However, prognostic prediction does not belong either on the classic learning paradigms of function approximation or classification. This is due to a patient can be classified as a "recur" case (instance) if the disease is observed, while there is no a threshold point at which the patient can be considered as a "non-recur" case. The data are therefore censored since a time to recur for only a subset of patients is known. For the others patients, the length of time after treatment during which malignant masses are not found is

known. This time interval is the disease free survival (DFS) time, which can be reported for an individual patient or for a study population. In particular, the right endpoints of the recurrence time intervals are right censored, as some patients will inevitably change hospital, doctors or die of other unrelated with the cancer causes. Therefore, the training dataset for the learning phase is not well-defined. Several groups have approached prognosis as a separation problem using different learning architectures such as back propagation artificial neural networks [14], entropy maximization networks [15], [1] decision trees [17] and fuzzy-based measurements [18].

In this paper is proposed a innovative approach to automatically detect the breast cancer using Fuzzy C-Means data mining techniques. This approach utilizes fuzzy c-means clustering for classification of the data from the WBCD dataset. The rest of paper is organized as follows. Section 2 describes data set of breast cancer disease. Section 3 presents of fuzzy c-means algorithm for classification. Section 4 presents shows results and finally sections 5 conclude the paper.

2. DATA SET OF BREAST CANCER DISEASE

The WDBC and WPBC datasets are made at the University of Wisconsin Hospital for the diagnosis and prognosis of breast tumours solely based on FNA test. This test involves fluid mining from a breast mass using a small-gauge needle and then visual checkup of the fluid under a microscope. This dataset is created by Dr. William H. Wolberg from University of Wisconsin Hospitals.

Table1. Attributes and values of cancer clinical instances

No	Attribute	Values
1	Sample code number	Id number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class	2 and 4

This dataset contains a total of 699 clinical cases, with 458 benign and 241 malignant cases. Each and every clinical case has 9 attributes with assigned integer values varying from 1 to 10 and one class output with a binary value of either 2 or 4, showing benign and malignant breast cancer diagnoses, separately. The table 1 showed the physical meaning of the nine attributes. Among the 699 clinical cases, 16 instances are each missing one of the nine attributes. The dataset consists of 683 cases, with each entry indicating the classification for a certain ensemble of measured values. For a consistently of high accuracy, the 16 cases each have missing one attribute are removed from this dataset. The resulting dataset has 683 clinical cases, with 444 (65.01%) benign and 239 (34.99%) malignant diagnoses. The evolutionary experiments executed fall into two categories, according with the data repartitioning into two dissimilar sets: training set and test set. The experimental categories classified are: (1) data set which contains all 683 cases of the WBCD dataset (2) training set that contains 70 cases

3. PROPOSED FUZZY C MEANS METHOD

Complexity of medical diagnosis problems has showed that using traditional methods in solving these issues is not appropriate. In medicine, the absence of information, and its roughness, and contradictory nature is common facts. Fuzzy logic plays an important part of diagnosis the medical disease. Some examples showing that fuzzy logic involving many disease groups are the following [19]:

(1) To analyze diabetic neuropathy
(2) To determine appropriate lithium dosage
(3) To calculate volumes of brain tissue from magnetic resonance imaging
(4) To characterize stroke subtypes and coexisting causes of ischemic stroke.
(5) To improve decision-making in radiation therapy
(6) To control hypertension during anesthesia
(7) To determine flexor-tendon repair techniques
(8) To detect breast cancer, lung cancer
(9) To assist the diagnosis of central nervous systems tumors (Astrocytes tumors)
(10) To discriminate benign skin lesions from malignant melanomas
(11) To visualize nerve fibers in the human brain
(12) To represent quantitative estimates of drug use
(13) To study the auditory P50 component in schizophrenia

3.1 Fuzzy c-means algorithm

Clustering is a process of grouping data in clusters, where data placed in one cluster are more similar to each other than those in other clusters. Fuzzy c-means algorithm uses the reciprocal of distances to decide the cluster centers. Fuzzy logic was introduced by Zadeh during 1960s for handling the uncertain and imprecise knowledge in real world applications [20]. Fuzzy C Means centroid of a cluster is calculate as mean of all points' value, weighted by their degree of belonging to the cluster. Advantages of this algorithm are that this method gives better results than k-means algorithm. Furthermore the greatest advantage of using fuzzy logic lies in the fact that scientists can model complex systems by implementing human experience, knowledge, non-linear, and imprecise and practice as a set of inference (or fuzzy) rules that use linguistic or fuzzy variables [21]. The FCM algorithm is to improve the accuracy of clustering under noise. FCM method is created on minimization of the function:

$$J_{m=\sum_{i=1}^{N}\sum_{j=1}^{C} uij^{m}\|Xi-Cj\|^{2}, 1\le m\le\infty} \quad (1)$$

Where u_{ij} represents degree of membership element x_i in the cluster j. The squared element is the Euclidian distance between i[th] data and j[th] center of cluster. Completed every iteration update of the membership function and center of clusters c_j is calculated as following:

$$uij = \frac{1}{\sum_{k=1}^{C} \frac{(\|Xi-Cj\|)^{\frac{2}{m-1}}}{\|Xi-Ck\|}} \quad (2)$$

Where centers of the clusters can be calculated as follows:

$$Cj = \frac{\sum_{i=1}^{N} uij^{m}}{\sum_{i=1}^{N} Uij^{m}} \quad (3)$$

Algorithm performs calculation as follows:

1. Initialization of *U=[uij], U(0)*
2. Calculation of centers of the vectors *C(k)=[cj] and U(k)*
3. Update of *U(k) to U(k+1)*
4. Comparison, if absolute value ‖U(k+1)-U(k) ‖ < ε, where ε represents predefined criteria, then ST In this study, 683 clinical instances in the Breast Cancer

4. RESULT

Wisconsin (Original) Dataset were used for this model. Even original data from has 699 clinical cases 16 cases were removed because of missing of one or more attributes. In rest 683 clinical cases there are 444 benign which represent 65% and 239 malignant breast cancer cases which represent 35%. The class output of an original binary value was 2 or 4 indicating benign and malignant breast cancer, one to one. Estimation of error of this model is done using two approaches such as training set and test set.

In Table1 are shown results using this method. In Fig.1 is shown scheme of the model and steps performed during the evaluation of the results.

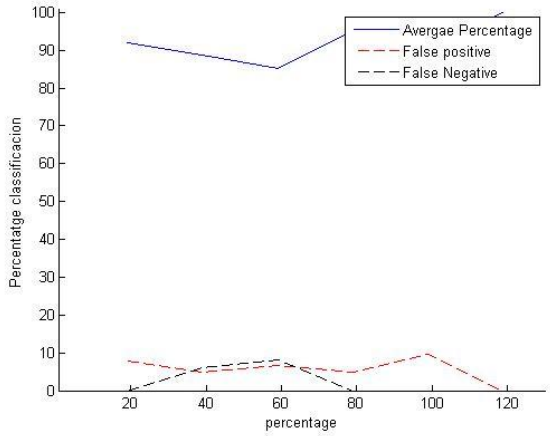


Fig 1. Evaluation Results of WPBC Dataset

Table 1: WPBC Data set result

Data	FCM Classification	Disease Diagnosis
True positive	100 %	100 %
True negative	87%	80.5%
False positive	0 %	0 %
False negative	13%	19.5%

In Table 2 are shown results using this method. Fuzzy c-means is done with initial data set as well as training data.

In this approach, WBCD data set divided to 10 parties, nine parties for train set and one party for test set. This process runs for more than ten times. Results of these run is indexed in table 2. Average of train set accuracy and test set accuracy is 98.91 and 97.99, in sequence.

Proposed approach is compared with some algorithms, such as Naïve Bayes, SVM, MLP. Result of comparison is indexed in table 3. According to Table 3, train set accuracy of FUZZY-C means algorithm (proposed algorithm) better than other algorithms and also test set accuracy of FUZZY-ACO algorithm better than other algorithms. However, the main advantage of proposed algorithm is high interpretability.

Table 2. Results of proposed approach

Run	1	2	3	4	5	6	7	8	9	10
Train	614 /629	614 /629	619 /629	613 /629	622 /629	624 /629	615 /629	619 /629	620 /629	616/630
Test	68 /70	69 /70	67 /70	69 /70	67 /70	69 /70	68 /70	68 /70	67 /70	66/69
Rules	20	20	20	20	20	20	20	20	20	20
Length	1.3	1.4	1.3	1.1	1.55	1.05	1.2	1.4	1.3	0.75

Table 3. Results Comparison of some algorithms

Algorithms	*MLP*	SVM	Naïve Bayes	Fuzzy C Means
Test Set	94.56	96.99	96.56	97.13
Train Set	97.29	98.09	98.41	98.62

5. CONCLUSION

Breast cancer has become the leading reasons of death in women in most of countries. We need most effective techniques to reduce breast cancer a death is detect it earlier. Early diagnosis requires a reliable and accurate diagnosis procedure that allows physicians to decide benign breast tumors from malignant ones without going for surgical biopsy. In this paper, is proposed a new alternative approach for breast cancer disease diagnosis and classifying benign and malignant breast cancer using fuzzy c-means. This proposed approach was based classification of input data, training data and test data. Results on breast cancer diagnosis data set from UCI machine learning repository show that the proposed FUZZY C Means would be capable of classifying cancer cases with high accuracy rate in addition to adequate interpretability of extracted rules.

7. REFERENCES

[1] G. Salama, M.B. Abdelhalim, and Magdy Abd-elghany Zeid. Son, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-classifiers*" International Journal of Computer and Information Technology*, vol. 01, pp. 36-43, September 2012.

[2] George J.Miao, Kathleen H.Miao, Julia H.Miao, "Neural pattern Recognition Model for Breast Cancer Diagnosis*" Journal of selected areas in Bioinformatics, August edition,2012,*pp. 1-8.

[3] U.S. National Institutes of Health, National Cancer Institute, http://cancernet.nci.nih.gov/

[4] Wang, T.C., Karayiannis N.B., Detection of microcalcifications in digital mammograms using wavelets, IEEE Transactions on Medical Imaging, Vol. 17, Issue. 4, Aug. 1998, pp. 498 – 509.

[5] Huo, Z., Giger, M., Vyborny, C., Wolverton, D., Schmidt, R., Doi, K., Automated computerized classification of malignant and benign mass lesions on digital mammograms, Acad. Radiol. 5, 155–168, 1998.

[6] Cheng Heng-Da, Lui Yui Man, Freimanis R.I., IEEE Transactions on Medical Imaging, Vol. 17, Issue. 3, June 1998, pp. 442 – 450.

[7] Pendharkar P.C., Rodger J.A., Yaverbaum G.J., Herman N. and Benner M., Association, statistical, mathematical and neural approaches for mining breast cancer patterns, Expert Systems with Applications, 17:223–232, 1999.

[8] Setiono R., Generating concise and accurate classification rules for breast cancer diagnosis, Artificial Intelligence in Medicine, 18:205–219, 2000.

[9] Chen D., Chang R.F., Huang Y.L., Breast cancer diagnosis using self-organizing map for sonography, Ultrasound in Medical Biology 2000, Vol. 26, pp. 405–11.

[10] Giger M., Huo Z., Kupinski M., Vyborny C., Computer-aided diagnosis in mammography. In Handbook of Medical Imaging, (Eds.) Sonka, M., Fitzpatrick, J., Medical Image Processing and Analysis, Vol. 2. SPIE Press, pp. 917–986, 2000.

[11] Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E., A neural network approach to breast cancer diagnosis as a constraint satisfaction problem, Med. Phys. Vol.28, pp. 804–811, 2001.

[12] S.Saheb Basha, Satya Prasad, "Automatic detection of breast cancer mass in mammograms using morphological operators and fuzzy c-means clustering" *Journal of theoretical and applied information technology*, pp. 704-709.

[13] http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/, Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset.

[14] Burke H. B., Goodman P.H., et al, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer, Vol. 79, pp. 857-862, 1997.

[15] Choong P.L, deSilva C.J.S et al., Entropy maximization networks, An application to breast cancer prognosis, IEEE Transactions on Neural Networks, 1996, 7(3):568-577.

[16] Choong P.L., deSilva C.J.S, Maximum entropy estimation vs. multivariate logistic regression: which should be used for the analysis of small binary outcome data sets?, Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol.3, pp:1602 – 1605, 1998.

[17] Wolberg W.H., Street W.N., Heisey D.M., and Mangasarian O.L., Computer-derived nuclear features distinguish malignant from benign breast cytology, Human Pathology, 26:792--796, 1995.

[18] Seker H., Odetayo M., Petrovic D., Naguib R.N.G., Bartoli C., Alasio L., Lakshmi M.S., Sherbet G.V., A fuzzy measurement-based assessment of breast cancer prognostic markers, Proceedings of the 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 9-10 Nov. 2000, pp.174 – 178.

[19] Angela Torres, Juan Nieto,"Fuzzy ogic in bioinformatics and medicine", Journal of Biomedicine and Biotechnology, Vol. 2006, pp. 1–7.

[20] Timothy J.Ross, "Fuzzy Logic with engineering applications", Third edition, Wiley, 2010.

[21] Victor Balanica, Ioan Dumitrache, "Evolution of breast cancer risk by using fuzzy logic" U.P.B.Sci.Bull, Vol. 73, 2011 pp. 54-64

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 618 - 622, 2013, ISSN: 2319–8656

Dissemination of Messages to Distant Vehicles in Sparse Condition via RSU as a Backbone

Jayanthi.S
Angel College of Engg. and Technology
Tirupur, TamilNadu, India

Rajeswari.M
Dept Of CSE
Angel College of Engg. and Technology
Tirupur, TamilNadu, India

Umamaheswari.P
Dept Of CSE
Info Institute Of Technology
Coimbatore, TamilNadu, India

Praveena.R
Angel College of Engg. and Technology
Tirupur, TamilNadu, India

Aswiga.R.V
Angel College of Engg. and Tech.
Tirupur, TamilNadu, India

Abstract: Vehicular Ad hoc NETworks (VANETs) are a variety of Mobile Ad hoc NETwork (MANET) that are used for communication among vehicles and between vehicles and road side equipments. For efficient communication, the routing protocols must be reliable and robust. In our work we deploy the RoadSide Units (RSUs) as an infrastructure and make use of it for the packet delivery. We carry and forward the message via RSU. In existing work they made use of RSU but that led to overhead and it is a time consuming, since they communicate as: source vehicle to its RSU, between RSUs and then from destination RSU to destination vehicle. But our work helps the user to query the destination RSU and receive the reply by requesting source RSU. And this source RSU communicates intermediate RSUs to communicate to the destination RSU and replies. Here the RSUs are connected. So, this helps in reducing the overhead and time taken. In addition our system provides the security to the user and the information transferred by producing Cryptographic MIXed key (CMIX).

Keywords: Vehicular Ad hoc NETworks (VANETs); RoadSide Units (RSUs); Reliable; Robust; Cryptographic mixed key (CMIX).

1. INTRODUCTION

The advances in the wireless technology paved way for the emergence of Vehicular Ad hoc NETworks (VANETs). As Mobile Ad hoc NETwork (MANET), VANET are also infrastructure less network but here the nodes participating are the vehicles.

VANET allow vehicles to connect between them. Different from MANET, VANET allows vehicles to move in an organized manner, since the restriction of roads, buildings, etc. The vehicular communication is classified as two sections: Vehicle to Vehicle communication (V2V) and Vehicle to Infrastructure (V2I) or Vehicle to Roadside (V2R).

Another major difference of VANET from MANET is that the routing protocols of MANET are not suitable for VANET in most scenarios. This is because VANET are formed by moving vehicles, have high node mobility and limited in mobility patterns. The major division of VANET routing protocols are: topology based and geographic based routing protocols. Many protocols are proposed under these categories only.

The vehicles in the VANET are equipped with On Board Unit (OBU), Global Positioning System (GPS), digital maps, navigation system, etc. The RSUs are deployed in various places with its coverage area. These RSUs are in turn connected with internet.

Our work is mainly concentrated on routing packets efficiently. Many proposed work has its own measures to provide good performance. Here we deploy many RSUs and make use of them to disseminate packets to distant vehicles. This works well in both dense and sparse conditions. Most of the proposed protocols works well in dense condition were the participating vehicles are available at most of the time. Whereas in sparse condition the participating vehicle density is low. Our work is motivated to get rid of this environment. To make the system work well in sparse condition, we deploy the RSU at certain intervals with their coverage range.

The system works as follows: The source vehicle S needs to sends the packet to the destination vehicle D. but the location of D is too far. At the time of routing, the density of vehicle is also less. So we make use of RSUs. The source vehicle s route the packet to its respective RSU. If there is available vehicle that directs to destination, then route the packet to this vehicle. Else the RSU disseminate to the neighboring RSU (may be destination vehicle's RSU) and from this intermediate RSU to other intermediate RSUs if needed or directs to the destination RSU. This RSU route the packet to the desired destination vehicle D.

Our work is different from the existing is: In existing work the source vehicle S needs to send a packet to the destination vehicle D which is also too far. There are RSUs at various regions. The source packet is sent to the RSU within its coverage. This RSU seeks for a vehicle to carry this packet and founds any then pass the packet. If no vehicles found then the packets resides in the RSU. This intermediate vehicle carries the packet and delivers to its RSU and so on. When the packet reaches the destination RSU, it is delivered to the destination vehicle. Here at each stage of forwarding the vehicles are involved. There is no direct contact between RSUs. So here the involvement of the vehicle takes time because the RSU waits for the vehicle and the time taken by this vehicle to carry the packet to next RSU. Also produce the overhead at RSU since the packet resides there still the desired vehicle is found.

So, our system gains the performance than the proposed and reduces the overhead and the time taken for forwarding.

The rest of this paper is designed as follows: the related work is depicted in section 2. In section 3 the proposed work is described. Finally section 4 the work is concluded and the future work is discussed.

2. RELATED WORK

Y. Ding et. al. proposed to improve the delivery performance by deploying static nodes at intersection. In addition the adjacent nodes measure the delay of forwarding data between each other in real time, so that the routing decision can adapt to changing vehicle densities.

SADV reduces the data delivery delay through three mechanisms: *SNAR (Static Node Assisted Routing)*: Here the intersection stores the path and forwards the best path.

LDU (Link Delay Update): Here the static nodes measure the link delay between each other in real time.

MPDD (MultiPath Data Delivery): It is used to decrease the packet delivery delay by trying to hit a faster delivery path.

C. Lochert et. al. proposed the Geographic Source Routing (GSR). This combines position based routing with the topological knowledge that is suited for city environments. Then they compared GSR with the other topology based approaches like DSR and AODV and conclude that GSR performs is enhanced than the other.

J. Zhao et. al. they concentrated on the problem of delay tolerant applications in the sparse network. This use the carry and forward method with the predicted mobility pattern. For this they projected some vehicle-assisted data delivery (VADD) protocols: L-VADD (Location First Probe), D-VADD (Direction First Probe), MD-VADD (Multipath Direction First Probe) and H-VADD (Hybrid Probe). This can be selected based on the techniques used for road selection at the intersection.

R. Frank et. al. they used the TrafRoute routing protocol that involves self-election and this self election is based on position, knowledge of the road topology and node density. Then this is suitable for both the V2V and V2I. This is very effective for short distance and to make use of the presence of infrastructure for longer distances.

R. Lu et. al. proposed the FLIP: An efficient privacy preserving protocol for finding like-minded vehicles on the road. This uses a security scheme Interest Privacy (IP). In this find the likeminded vehicles and secure sharing takes place via shared session key and distinguishes from those vehicles that are not like-minded.

3. PROPOSED WORK

In all above mentioned works they dint concentrate about the intersection issues. Because at the intersection the vehicle may take another route that never reach the destination vehicle. They just relied on moving vehicles to carry and forward packets.

But in our work we deploy RSUs on road side and at intersection, when needed. This RSU stores the packet and deliver to intermediate vehicle or to intermediate RSU. If intermediate vehicle change their path, then with the help of RSU we can request that packet to be resend. But automatically the RSU resend the failed packets if it dint receive any acknowledgement. So this helps in efficient routing without dropping any packets.

Then the security which plays a major role in VANET. For security we use Cryptographic MIXed key (CMIX), were cryptographic keys are generated for the users. In VANET for the vehicle to participate in communication, it must be authenticated and get certified by Certification Authority (CA).

Once the CMIX are generated then they use only those keys to authenticate themselves to other users. Each RSU authenticates the vehicles under its coverage. Only after authenticated by RSU the vehicle can participate in the communication. So this enables the robust routing.

3.1Proposed Architecture

The above described work is depicted by architecture diagram as in figure 1. In the figure, there are vehicles that are moving. They are authenticated by RSU to participate in the network. The RSUs are connected to the server.

The source vehicle S needs to send a packet to the destination vehicle the arrow 2.D. But the location of D is extreme. So S route the packet to the RSU of its region. This is indicated by the arrow 1. Next this RSU, communicate with the neighboring RSU and disseminate the packet. This RSU act as an intermediate node. This process is specified by the arrow 2. Then this RSU send the packet to its neighboring RSU which is the destination vehicles' RSU. This process is noted by the arrow 3. Finally from this RSU searches the destination vehicles' location and the packet is delivered to the destination vehicle. This process is notified by the arrow 4.

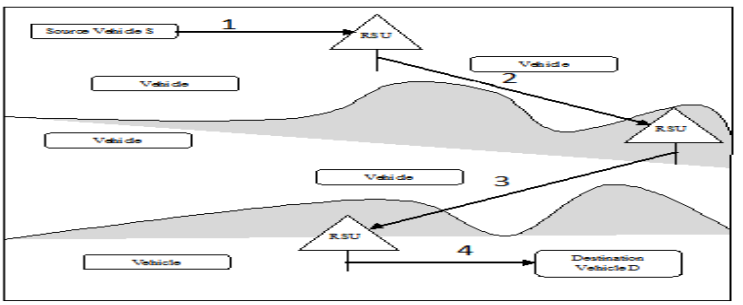


Figure 1. Proposed Architecture

4. CONCLUSION

In this paper we have projected the routing efficiency and robustness. The proposed work uses the infrastructure (RSU) as a backbone to route the packets to distant vehicles in sparse environment. This helps in reducing the load and complexity in transferring the data efficiently. And with this the security is ensured by CMIX that excludes the unauthorized vehicle.

5. References

[1]. R. Frank, E. Giordano, and M. Gerla, "TrafRoute: A different approach routing in vehicular networks," in *Proc. VECON, Niagara Falls, ON, Canada*, pp. 521–528, 2010.

[2]. J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," *IEEE Trans. Veh. Technol., vol. 57, no. 3, pp. 1910*–1922, May 2008.

[3]. Y. Ding, C. Wang, and L. Xiao, "A static-node assisted adaptive routing protocol in vehicular networks," in *Proc. VANET, New York*, pp. 59–68, *Sep. 2007*.

[4]. C. Lochert, M. Mauve, H. Füÿler, and H. Hartenstein, "Geographic routing in city scenarios," *SIGMOBILE, vol. 9, no. 1, pp. 69–72, Jan. 2005.*

[5]. R. Lu, X. Lin, X. Liang, and X. Shen, "FLIP: An efficient privacy preserving protocol for finding like-minded vehicles on the road," in *Proc. Globecom, pp. 1–5, 2010.*

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 623 - 627, 2013, ISSN: 2319–8656

Survey on Image Enhancement Techniques

P.Suganya
Vivekanandha College of Engineering for Women,
Tiruchengode,
Namakkal-637205
Tamilnadu, India.

S.Gayathri
Vivekanandha College of Engineering for Women,
Tiruchengode,
Namakkal-637205
Tamilnadu, India.

N.Mohanapriya
Vivekanandha College of Engineering for Women
Tiruchengode,
Namakkal-637 205
Tamilnadu, India.

Abstract: Enhancement is one of the challenging factors in image processing. The objective of enhancement is to improve the structural appearance of an image without any degradation in the input image. The enhancement techniques make the identification of key features easier by removing noise and other artifacts in an image. This paper analyzes the performance of various enhancement techniques based on noise ratio, time delay and quality. It also suggest suitable algorithm for remote sensing images based on the analysis.

Keywords: Image Enhancement, Histogram Equalization, Stochastic Resonance, Contrast Enhancement, Spatial domain, Frequency Domain and Noise ratio.

1. INTRODUCTION

Image Processing is a processing of image and takes image as an input, the output of image processing may be either an image or set of characteristics. This includes image enhancement, noise removal, restoration, feature detection, compression, etc. Digital images are always affected by noise, blurring, incorrect color balance and poor contrast. Most of digital images that can be produced through scanners, digital cameras, video cameras, Charged Coupled Devices (CCD cameras) and web-cam can be easily affected by the these problems. This will lead to low quality images. Image enhancement will be used to minimize the effects of these degradations. This can be done by using a number of image enhancement techniques. Specifically, an enhancement of color image is to process the luminance and color information to make an image has sharp details, rich in color and better visual effect without any distorting or shifting of color. The image enhancement is to process an image so that the result is more suitable than the original image for specific application. The enhancement technique applied for various applications such as medical images, remote sensing images and general images. The objective is to improve the characteristic of an image to get clear image [13]. The enhancement methods can be broadly categorized into following two methods:

1. Spatial Domain Method
2. Frequency Domain Method

The spatial domain techniques, directly operates on pixels of an image. The pixel values are manipulated to achieve desired enhancement. The gain of spatial based domain technique is that they conceptually simple to understand and the complexity of these techniques are low [15]. But these techniques have difficult to providing sufficient robustness and imperceptibility requirements.

In frequency domain methods, the image is transferred into frequency domain. It means that, the Fourier transform of the image is computed first. The result of Fourier transform is multiplied with a filter transfer function. And then the inverse Fourier transform is performed to get the resultant image. Frequency domain image enhancement is used to describe the analysis of mathematical functions and signals with respect to frequency and operate directly on the transform coefficients of the image, such as Fourier transform, discrete wavelet transform (DWT), and discrete cosine transform (DCT). The advantages of frequency domain are, less computational complexity, manipulating the frequency composition of the image [11]. The disadvantages are, it cannot simultaneously enhance all parts of image in good manner and it is also difficult to automate the image enhancement procedure. Image enhancement is applied in every field where images are ought to be understood and analyzed, this section briefly describe the various image enhancement techniques.

Image enhancement means, transforming an image f into image g using T. The values of pixels in images f and g are denoted by r and s, respectively. As said, the pixel values r and s are related by the expression,

$$S = T(r) \qquad (1)$$

Where T is a transformation that maps a pixel value r into a pixel value s [1]. The results of this transformation are mapped into the grey scale range. So, the results are mapped back into the range [0, L-1], where L=2k, k being the number of bits in the image being considered. So, for instance, for an 8-bit image the range of pixel values will be [0, 255].

2. IMAGE ENHANCEMENT TECHNIQUES

The enhancement doesn't increase the inherent information content of the data, but it increases the dynamic range of the chosen features so that they can be detected easily. Few enhancement techniques are to be described below for color and gray scale images:

2.1 Histogram Equalization

Histogram of an image is concerned with the gray levels. Using histogram to decide that given image is whether a dark image or light image or low contrast or high contrast image. It can be expressed using discrete function as,

$$P(r_k) = \frac{n_k}{n} \qquad (2)$$

Where r_k denotes kth gray level, n_k denotes number of pixels in the image, n denotes total number of pixels and $k=0, 1, 2 \ldots\ldots 255$. Histogram Equalization which stretches histogram

to an image. It is used to improve the visual appearance of an image [10]. This technique involves,

1) Dividing image into segments.
2) Histogram is applied to find out the pixel intensity values for the gray levels and the image have gray levels or intensities in the range from 0 to 255.
3) Histogram Equalization is used to calculate the intensity values and make them uniform distribution of pixels to get an enhanced image.

Thus HE technique is used to increase the dynamic range of pixels for the appearance of an image.

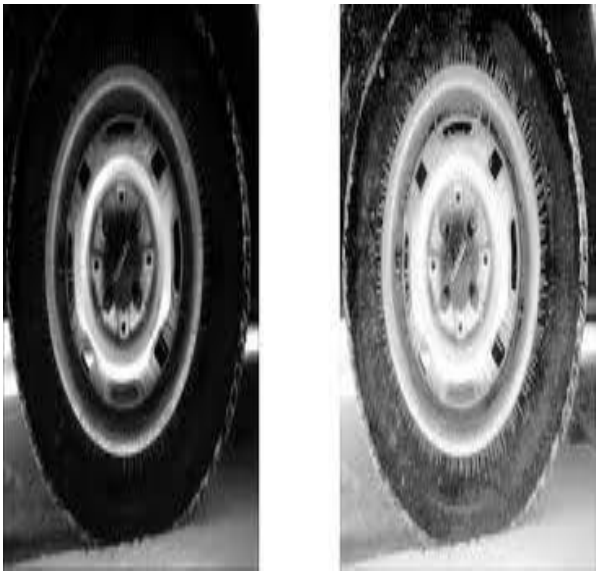
(a) (b)

Figure. 1 (a) Original Image (b) Enhanced image for Histogram Equalization

2.2 Brightness Preserving Bi-Histogram Equalization (BBHE)

The overall BBHE technique is used for preserving of brightness of an image. Brightness preservation is one of the most important characteristics of an image. So this method splits the image's histogram into two independently equalized parts. So the intensities are arranged equal as well. One drawback of the histogram equalization can be found on the fact that the brightness of an image can be changed after the histogram equalization, which is mainly due to the flattening property of the histogram equalization [3]. Thus, it is rarely utilized in consumer electronic products such as TV where preserving the original input brightness may be necessary in order not to introduce unnecessary visual deterioration.

The BBHE is extension of histogram equalization to overcome such a drawback of histogram equalization [7]. The essence of the algorithm is to utilize independent histogram equalizations separately over two subimages obtained by decomposing the input image based on its mean with a constraint that the resulting equalized subimages are bounded by each other around the input mean. It is shown that the proposed algorithm preserves the mean brightness of a given image significantly well compared to typical histogram equalization while enhancing the contrast and provides a typical enhancement that can be utilized in consumer electronic products. The output is shown below:



(a) (b)

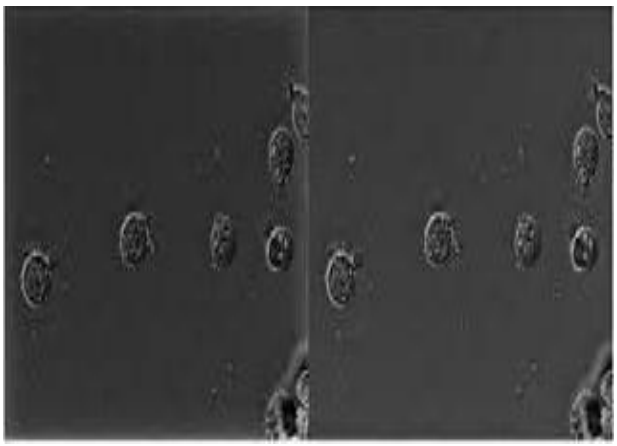
Figure. 2 (a) Original Image (b) Output Image for BBHE

2.3 Brightness Preserving Dynamic Histogram Equalization (BPDHE)

BPDHE is an extension of Histogram Equalization. In Dynamic Histogram Equalization (DHE) the input image's histogram is divided into partitions and so called sub-histograms. The DHE method is also used to provide mean brightness for an image and gives the intensities to have a new range [8]. It provides realistic images by look. In this method the intensities are equalized individually.

BPDHE is an extension to the DHE method. It shifts the mean brightness between the resultant histogram image and original image. So the mean brightness is preserved. And it produces the mean intensity of input and output images as equal.

The BPDHE technique uses different filters such as smoothing filter, Gaussian filter, etc. which smoothes the data by suppressing image noise for the clear image [9]. In addition to BBHE, DHE method provides better mean brightness for an image.

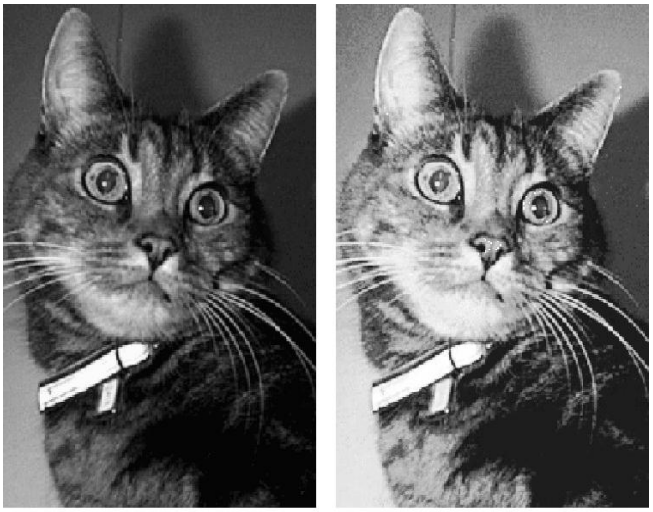


(a) (b)

Figure. 3 (a) Input image (b) Output Image for BPDHE

2.4 Adaptive Histogram Equalization (AHE)

Adaptive Histogram Equalization is used for improving contrast in images. It differs from Histogram Equalization by adaptive method that computes several histograms and each histogram corresponding to a distinct section of an image. The contrast of region for an image will not be sufficiently enhanced by Histogram Equalization. AHE improves this enhancement by transforming each pixel with a transformation function derived from a neighborhood region. It is used to overcome some limitations of global linear min-max windowing method. Thus it reduces the amount of noise in regions of the image. And also AHE have the ability for improving the contrast of grayscale and color image.

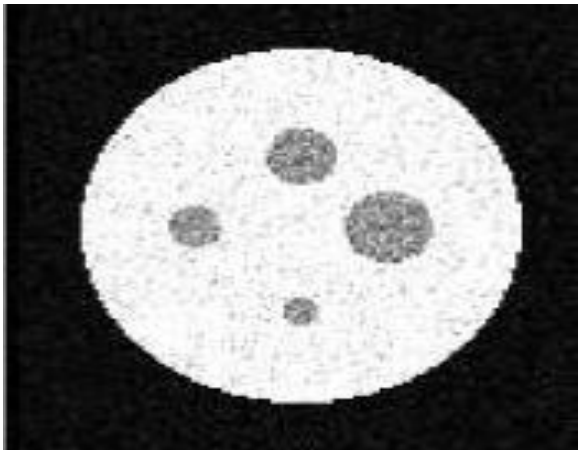


(a) (b)

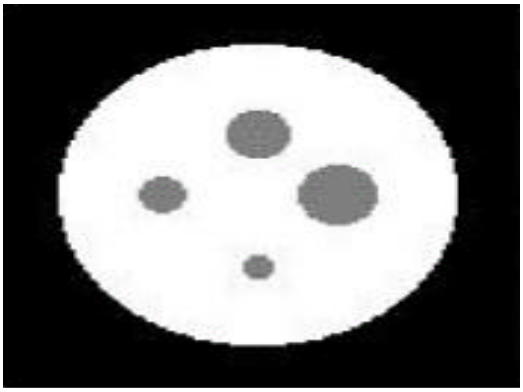
Figure .4 (a) Original image (b) Output Image for AHE

2.5 Stochastic Resonance(SR)

Stochastic resonance is broadly applied o describe any occurrence where the presence of noise in nonlinear system is beer for output signal quality then it absence [4]. To enhance the contrast of an image it utilizes external noise of an image.



(a)



(b)

Figure. 5 (a) Input (b) Output Image for SR

2.6 Contrast-Limited Adaptive Histogram Equalization (CLAHE)

To enhances the contrast of the grayscale image by transforming the values using contrast-limited adaptive histogram equalization (CLAHE).it operates on small regions in the image, called *tiles*, rather than the entire image [12]. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the histogram specified by the distribution parameter. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image.

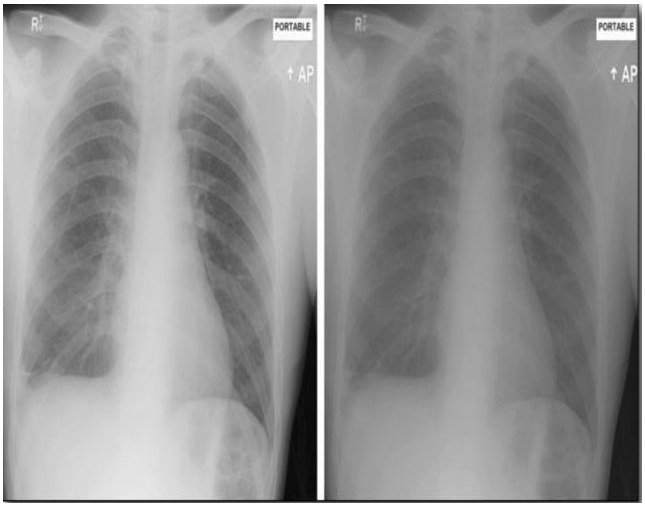
Figure 6. Original Image and Enhanced Image for CLAHE

2.7 Contrast Enhancement

This technique automatically brightens images that appear dark or unclear. Apply appropriate tone correction to deliver improved quality and clarity [2]. This play an important role in medical applications. This because of visual quality is very important to diagnosis diseases. X-Ray used to capture the internal structure of human body. It especially useful for check bone fracture. There are many advantages but X-Ray technology but it generates low contrast image due to presence of bulk amount of water in human body.

Image enhancement also perform automated X-Ray check system for making X-Ray images with more visual and contrast by using some contrast enhancement technique .Zooming an image an important task in many application .while zooming an image the pixels are inserted to enlarge the size of image.

The main task is interpolation of new pixel form surrounding the original pixel [6]. In weighted median used for edge preservation and less blocky look to edges. The Cathode Ray (CR) image of a patient's chest displayed with contrast enhancement on the left and unprocessed on the right for Contrast Enhancement is shown below using MATLAB.



(a) (b)

Figure. 7 (a). Original Image (b).Enhanced Image

2.8 Adaptive DWT based DSR

The DWT technique is used to produce high frequency content images. The DWT which decomposes the input image into sub bands. They are Low-Low (LL), Low-High (LH), High-Low (HL), and High-High (HH). The process of image using DWT is carried out by interpolating high-frequency sub band images and the low-resolution input images to produce the enhanced image [5]. The Adaptive DWT based DSR technique presented for perform enhancement of very dark images. It using inter noise to improve the performance of input image. It gives better enhancement for very dark images. It leads to less computational complexity [14]. This Technique is applied for enhancement of very dark images.

In Dynamic Stochastic Resonance (DSR) an external noise of an image is considered for an image. And the Adaptive DWT based Dynamic Stochastic Resonance uses internal noise for improving performance of an input image. It produces output without artifacts, ringing, blocking of the image. The adding of noise to the input image is useful for non-linear systems using this technique. By using lower noise intensities in SR mechanism the signal cannot be able to reach the threshold value. In this technique the noise allows the signal to reach the threshold value. Thus Adaptive DWT based Dynamic Stochastic Resonance is suitable for enhance both the grayscale and colored image.

3. PERFORMANCE ANALSIS

This paper collected various image enhancement techniques. In this section the performance of various image enhancement techniques have been specified in the below Table 1.

Table 1. Comparison of Enhancement Techniques

Enhancement Techniques	Advantage / Dis Advantage	Noise ratio	Time delay (ms)
Histogram Equalization	Preserves the background brightness / Not much suitable for color images.	24.3442	2.0
BBHE	Maintains the mean brightness / Takes more computational time.	25.1157	1.8
BPDHE	Produces intensity range of input and output images as equal / Does not give clear contrast.	24.4065	1.9
AHE	Contains low contrast with dark regions of image / Creates some unwanted blurring in edges.	30.2665	1.2
SR	Provides better signal quality for output image / Technique used for very low contrast image.	23.5472	1.6
CLAHE	Avoids amplifying noise that might present in image	30.7692	1.0
Contrast Enhancement	Gives clear contrast for X-Ray images / More computational requirement.	29.5928	2.0

4. CONCLUSION & FUTURE WORK

This paper have discussed about various enhancement techniques with their performance analysis using MATLAB tool with appropriate output shown in the above table. The output of each technique showed that improved image quality and better structural appearance of an image. And also increased dynamic range of pixels with better contrast, keeps the overall brightness level and the edges are preserved without any degradation. Even though all the techniques gave better result, the combination of Adaptive Histogram Equalization (AHE) and Contrast-Limited Adaptive Histogram Equalization (CLAHE) yields good performance for remote sensing applications. Because the AHE is contains low contrast with dark regions. The CLAHE technique better in contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image.

In future work, these enhancement techniques are to be applied for video images and 3D images.

5. REFERENCES

[1] Rafael C Gonzalez and Richard E Woods, "Digital Image Processing", third edition, Pearson Education, 2007.

[2] S.S. Bedi, Rati Khandelwal," Various Image Enhancement Techniques- A Critical Review", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 3, March 2013.

[3] Chao Wang and Zhongfu Ye "Brightness Preserving Histogram Equalization with Maximum Entropy: A Variational Perspective", Vol. 51, No. 4, November 2005.

[4] P. Hanggi, P. Jung, and F. Marchesoni, "Stochastic resonance", Rev. Mod. Phys., vol. 70, 223–270, 1998.

[5] Hasan Demirel and Gholamreza Anbarjafari, "Discrete Wavelet Transform-Based Satellite Image Resolution Enhancement" VOL. 49, NO. 6, JUNE 2011.

[6] Hassan, N. Y. and Aakamatsu, N., "Contrast Enhancement technique of dark blurred Image", International Journal of Computer Science and Network Security (IJCSNS), Vol. 6, No. 2, 2006, pp. 223- 226.

[7] Kim's, Min Chung, "Recursively Separate and Weighted Histogram Equalization for Brightness Preservation and Contrast Enhancement", IEEE Transaction on Communication, Networking and Broadcasting, Page: 1389-1397, Publication year: 2008.

[8] Kong.N.S.P, Ibrahim .H, "Color Image Enhancement using Brightness Preserving Dynamic Histogram Equalization", IEEE Transaction on Communication, Networking and Broadcasting, Page: 1962-1968, Publication year: 2008.

[9] Kuo-Liang Chung, Yu-Ren Lai, Chyou-Hwa Chen, Wei-Jen Yang, and Guei-Yin Lin, "Local Brightness Preservation for Dynamic Histogram Equalization", 2011.

[10] MandeepKaur, K iran Jain, Virender Lather International Journal of Advanced Research in Computer Science and Software Engineering "Study of Image Enhancement Techniques : A Review" Volume 3, Issue 4, April 2013.

[11] Nancy, Er. Sumandeep Kaur," Image Enhancement Techniques: A Selected Review", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 6 (Mar. - Apr. 2013), PP 84-88.

[12] Papiya Chakraborty, "Histogram Equalization by Cumulative Frequency Distribution", International Journal of Scientific and Research Publications, Volume 2, Issue 7, July 2012.

[13] Parth Bhatt, Sachin Patel", Image Enhancement Using Various Interpolation Methods", International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol. 2, No.4, August 2012.

[14] Rajlaxmi Chouhan, C. Pradeep Kumar, Rawnak Kumar, and Rajib Kumar Jha, "Contrast Enhancement of Dark Images using Stochastic Resonance in Wavelet Domain", International Journal of Machine Learning and Computing, Vol. 2, No. 5, October 2012.

[15] Ramkumar.M, Karthikeyan.B," A Survey on Image Enhancement Methods", International Journal of Engineering and Technology (IJET), ISSN: 0975-4024 Vol 5 No 2 Apr-May 2013, 960.

International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 628 - 636, 2013

Clogging Report with Ad-Diffusion and Carter Performance Inspection using Intelligent Vanet

Aswiga R.V.
Angel College of Engg. and Technology
Tirupur, TamilNadu, India

Rajeswari M.
Dept Of CSE
Angel College of Engg and Technology
Tirupur, TamilNadu, India

Umamaheswari P.
Dept Of CSE
Info Institute Of Technology
TamilNadu, India

Praveena R.
Angel College of Engg. and Techology
Tirupur, TamilNadu, India

Jayanthi S.
Angel College of Engg and Technology
Tirupur, TamilNadu, India

Abstract-Mobile Ad-hoc network is a collection of mobile nodes that form a temporary network without use of any existing infrastructure. Group communication is challenging task in MANET .In recent years many protocols have been developed to achieve group communication in MANET. Since VANET is the subset of MANET the same group communication would be achieved to improve the scalability, reliability and optimality of routing protocols. Using group communication, distributed traffic information system has evolved for detecting traffic problems on road. The traffic information will be communicated to vehicles in the form of V2V or V2I interaction medium. In the existing system, the traffic information with the vehicle id, position, location along with the ad dissemination, driver behavior information will be transmitted to other vehicle which leads to unnecessary tracking of vehicles on the road side. This in case leads to major threats such as attacks, hacking vehicles, etc...Therefore in proposed system the security in VANET is obtained by restricting the unnecessary flow of information like vehicle id or position to other ongoing vehicles on the roadside with the help of trust based reputation system in VANET. The ad dissemination along with traffic information and driver behavior information will be transmitted to other vehicles in the existing system with the help of trust based reputation system in VANET. But since it contains some disadvantages, the fuzzy logic reputation system will be implemented as our future work where the architecture is decentralized.

Keywords-*CEPA (Complex Eventual Parallel Architecture), CIS (Congestion Information System), IS (Information System), Agents, Vanet parallel ad dissemination architecture*

1. INTRODUCTION

With the rapid growth of traffic in internet, congestion control has become one of the most important issues in recent communication networks and in modern transmission technologies .But during the last decade's global road traffic has been increasing day by day. This leads to the fact that currently, road traffic congestion are one of the most common criteria that motorists have to tolerate in their trips. Changqiao Et al [4] illustrates that one of the most important topic in intelligent transportation system is the development of replicated congestion information systems (CIS).This system will identify the blockage conditions on the road and try to detect and avoid traffic flow difficulties through sequential delivery of messages between vehicles. Replicated congestion information system (CIS) will avoid the constraints on the road side and try to prevent the difficulties on the road side equipment approaches. Communication between vehicles is usually achieved with the help of ad-hoc network. So VANET is a movable dynamic network in which the nodes are designated as vehicles that travel on a Road and communicate with each other through wireless technology. Each vehicle in the network will broadcast data messages called beacons, which contains its current velocity, speed, position, etc…In this manner each vehicle of VANET can identify the neighboring vehicles that travel along the roadside. However these broadcasted messages imply a large number of events across the VANET that each vehicle must deal in the network. To cope with large number of events like accidents, traffic jam and other natural disasters, an eventual architecture has been developed as a new architectural paradigm. Eventual architecture is a software model which processes the events with low delay where the events should be processed as soon as they arrived. Currently eventual architecture plays an important role in many e-commerce areas such as financing or goods distribution. Furthermore this is expected to grow in future as long information systems need to evaluate data from many more replicated sources. In an E-commerce domain a certain information systems are used and frequent real world incidents on the road are designated as events in first layer of information systems (IS).These incidents have complex relationships that contain similar patterns. Real time incidents here indicate accidents, traffic jam and other natural disasters that occur on the road. Complex relationship deals with vehicles that are travelling below the average speed on the road and group these vehicles into a cluster. Because of processing these real time incidents in the eventual architecture, it is possible to become aware of target real world activity by checking the information systems and looking for these similar patterns. Consequently the key task is the subsequent identification of particular event pattern from the stream of events that flow through IS. This is the sophisticated form of software architecture that deals with numerous heterogeneous events from different streams. This type of system is suitable and adoptable for replicated environments where dispersed elements can act as data a source which provides useful information about the environment in which vehicles travel along the road side. In this paper we proposed an architecture in such a way that accepts the data's from different sources but process only beacon messages in order to detect the traffic or blockage conditions on the road. Other messages will not be processed inside the

architecture except the beacon messages. The input for beacon messages from different sources web sources, on-board sensors, vehicles, etc. This method helps to identify the traffic conditions and differentiate it based on different lanes regardless of the communication protocol used. This takes real world map thus covering large amount of data's from road side, This paper takes an initial approach to filter only the traffic messages along with ad disseminations and identify whether the driver is alive or not and process the information containing the corrective action to other vehicles .Fernando et al [1] demonstrated that the security is enhanced by restricting the flow of information like vehicle id, location, etc to other vehicles in order to avoid major threats like terrorist attack, robbery, etc..Parallel traffic management system will be implemented as future work.

2. RELATED WORK

Vehicular ad-hoc network has widely been used to develop traffic information system (TIS). This traffic system gathers information from messages to check the traffic conditions on the road. This reduces the accidents by alerting motorists who is travelling on the congested location. Based on this we propose a system called as a replicated congested information system (CIS).In this system each vehicle communicates only the traffic messages to other vehicles in a secure way. In this approach we extend the architecture as parallel architecture that guarantees the following different characteristics like choosing the alternative path to reach the destination and control the traffic with the help of security measures for various transportation services which includes public vehicles, taxies, metro, fire trucks, ambulances, etc…Fluctuation of traffic status is very difficult in order to predict and control the traffic messages. In addition to the existing work we also included the ad dissemination along with the driver's behavior detection system that transmits the corrective action to the current vehicle and also to other ongoing vehicles on the road side. Thus in this paper we reproduced the architecture in such a way that it allows dissemination of traffic information along with the ad dissemination and produces safety measures that guides other vehicle driver's to take corrective action from the existing information provided .

3. ARCHITECTURE OVERVIEW

EDA act as a middleware between network layer and back-end application. But we are adding an additional parallel system next to EDA which acts parallel and provide decisions to be made in case of any accidents or some natural disasters on the roadside .This is called as CEPA (Complex Eventual Parallel Architecture). This architecture will act parallel by taking into account the input from different sources like EDA, weather conditions, external source agent, etc.., On the other hand, network layer handles the dissemination and reception of traffic messages throughout the VANET. Our proposed architecture also provides an interface that allows external entities to access information from underlying VANET in a quite efficient way. This CEPA generates a traffic alarm whenever a traffic jam is detected. This type of event is sent to the back end application. So that they could use this information to alert both the driver and the information panels of the motorway

depending on where CEPA is running. In Saif Al-sultan et al [3] the driver's behavior is detected with the help of sensors which is situated inside the vehicles. This will be monitored on the parallel system in order to take decision and send the message to other ongoing vehicles on the roadside to prevent accidents from happening. This parallel architecture will detect the abnormal behaviors which is exhibited by drivers and based on that evaluate the driver's condition while travelling on the road.CEPA generates different agents like traffic filter agent (TFA) to filter the traffic related information and identify the slow vehicles which is travelling on the road with the help of SVA (slow vehicle agent).SVA will detect the vehicles which are running below the average speed or even stopped in a busy environment. The vehicles will usually slow or stop flowing because of traffic or blockage which can be identified by SVA.Warning agent (WA) will provide warning messages to other vehicles after classifying the traffic level such as slight, moderate and severe. In addition to the above agents, the adaptation agent (AA) and external source agent (ESA) helps to take intelligent decisions. In previous work [20] we did not identify the decrease in diffusion rate of information when it reaches the far vehicles. But here we identify the diffusion rate and overcome this problem by designating each RSU to be a SRSU (source road side units). That is when the message needs to be delivered to the far vehicles and when the destination is in large distance; the intermediate relays can act as source hence the transmission is done with the help of logical transmission via signals. Since the signal is reached from the source which is nearer to the destination vehicle, the message will be delivered to the vehicle as designated by the sender. Here the malicious nodes will not transmit wrong information to vehicles because trust based reputation system is used here. The figure one demonstrates the work flow process of vanet reputation system architecture.

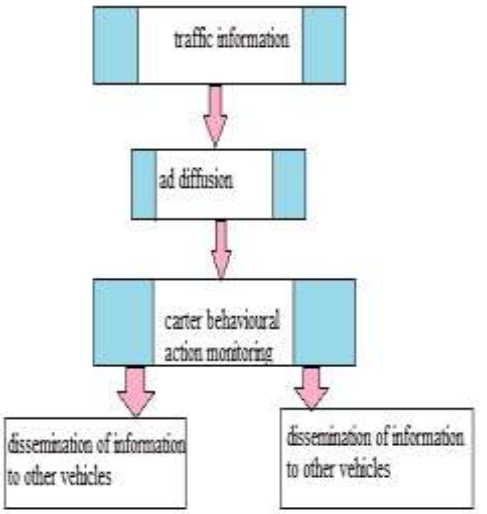


Fig.1. Work Flow Process Diagram

In the Fig.3 shown below demonstrates that, from the network layer different messages will be transmitted to the complex EDA. But this architecture will filter only the traffic messages with the help of TFA and identify the slow moving vehicles that are below the average speed .After getting input from various external source agent , if the blockage is detected then the adaptation agent will take corrective action with the help of experts knowledge and pass this information to parallel architecture which evaluates drivers behavior and calculate the number of vehicles flow rate for that particular location to help the vehicle to reach the destination in short time. Finally warning agent will send the

alarm to other ongoing vehicles on the road by restricting the traffic id in order to avoid security threats. As shown in this figure below control unit 1 passes only ad dissemination details. Control unit 2 passes driver behavior detection system information to other ongoing vehicles on the road side. Control unit 3 passes only warning message regarding the traffic information to other ongoing vehicles on the road side.

4. AD-DIFFUSION RATE

An SP refers to a business entity with a fixed position, such as a restaurant or a gas station. Each node periodically broadcasts beacons containing the driving states, such as location, speed, and heading direction, to support traffic safety applications. First, each SP intends to maximize its Advertising effect by disseminating ads to as many nodes as possible. Second, as an ad receiver, each node would like to learn of the local services without being distracted by excessive ads. In addition, being selfish, each node forwarding one ad expects to receive certain incentive in return. Third, VANETs as a whole need to ensure ad dissemination is under control in the face of increasingly more ads to avoid message storms. Since the information is delivered in the logical manner from the relay which acts as a source, the quality of message is high. The quality here is considered to be reliable. Reliable in the sense the message is not affected by malicious nodes or the attackers cannot change the content of message. Here the first vehicle transfers the information to other vehicles which act as a source unit that contains Fuzzy reputation system in vanet. The red color nodes obtain information from other vehicles but it act as a source when it delivers the message to other vehicles. The node stops acting as a source when it designates to be a final node.

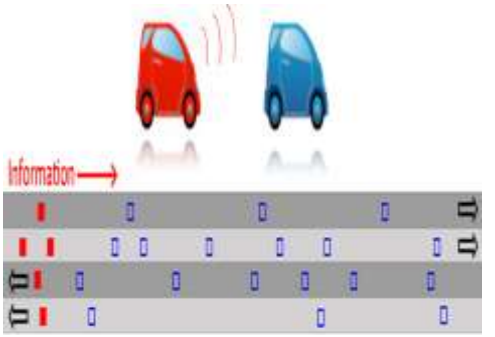


Fig.2. Clogging Report with Ad-Diffusion

5. VANET PARALLEL REPUTATION SYSTEM ARCHITECTURE IMPLEMENTATION

AD is proposed to support secure ad dissemination with pragmatic cost and effect control. To trade off the conflicting requirements of the involved parties, an incentive-centered architecture is proposed for VAAD, where the SP pays other entities for their services in ad dissemination. Constrained by the incurred cost, each SP will set a realistic advertising effect requirement in terms of the number of ad receivers and the ad rebroadcast frequency. The VAAD Manager (VM) is introduced to coordinate the interactions between SPs and VANETs. Upon receiving a dissemination request from one SP with cost and effect specifications, the VM will obtain proper authorization from VANET Authority for this ad. With the authorization, the VM can request one SRSU to disseminate the ad according to the specifications. The goal of trust management in VANETs is not limited

to reliable package delivery. One main aim of VANETs is to increase road safety and reduce traffic congestion by allowing information sharing among peers about road and traffic conditions. Trust management in VANETs should help peers detect false information provided by malicious nodes and make informed driving decisions. Trust management in this case is more challenging than that for reliable package delivery. Much dynamics has to be taken into consideration, such as the time and location of reported events, and the types of the events. Thus, previous trust modeling endeavors in mobile Adhoc networks become worthless when being directly applied
to vehicular ad-hoc networks. Trust management can effectively improve peer collaboration in VANETs to share information and detect malicious peers. However, the trust management itself may become the target of attacks and be compromised. Thus fuzzy logic based reputation system will be considered as our future work.

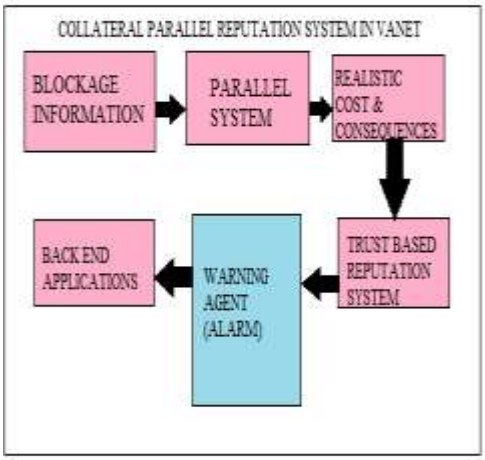


Fig.3.Vanet Parallel Reputation system Architecture implementation

6. SIMULAITON STUDY

6.1. Performance Evaluation

The simulation results bring out some important characteristic function of Region based clustering mechanism. In this section we record the Cost and Overhead Comparisons and packet ratio parameters of the simulation by using record procedure. The parameters includes Packet delivery ratio, Throughput, Packet loss rate, delay time etc. The recorded events are stored in the trace files. The number nodes is going to be participated in the simulation is decided. As it is a VANET environment, the topology changes dynamically. Here we are using only a logical topology as it is wireless environment. By executing the trace files by using xgraph or gnuplot we can get the graph as the output.

6.2. *Simulation Parameters*

With regard to the evaluation of the CEPA, the following two different measurements have been used: 1) the Evaluated/Detected rate (DR) and 2) the mean time to detection (MTTD). Both Measurements are defined by the following equations:

Detected Rate =Total No of detected traffic

Number of traffic jams On the Lane

$MTTD = 1n\_nI = 1(tdetection - tstart)$

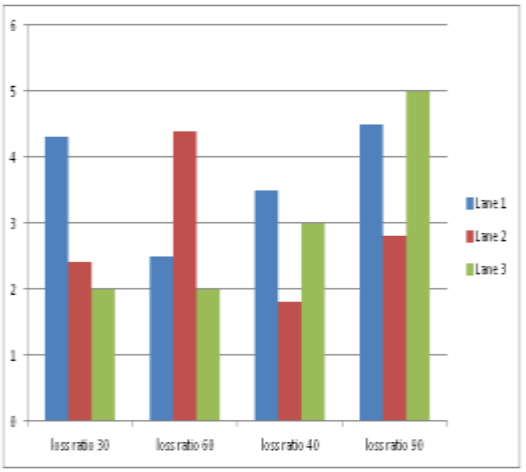


Fig. 4. Estimation rate of the CEPA for the two types of simulated congestions, given different penetration at y-axis and packet-loss rates at x-axis.

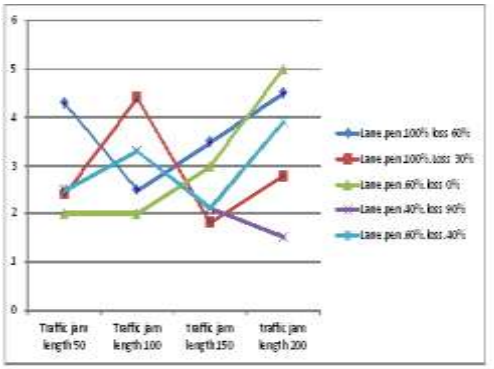


Fig. 5. MTTD of the *lane* and *raw* modes, given some penetration rates (60%, 40%, 100%) and loss ratios (60%, 30%, 40%, 100%) for a three-lane-type traffic jam.

7. CONCLUSION

In this paper, a complex eventual parallel architecture along with Ad dissemination has been put forward as a method for identifying traffic congestions in the context of a replicated CIS. In particular, this architecture may run as part of onboard equipment in a vehicle or in an external module of the road infrastructure. The system processes the beacon messages, and it can detect different levels of congestion on a road; moreover, it can add environmental information such as the current weather condition plus the behavior of the driver so that the traffic jam detection is enriched. Results from the different simulations state that this architecture can detect several types of traffic jams. Thus, the vanet parallel ad dissemination architecture achieved better results, as long as the traffic jam covered several lanes. Therefore in proposed system the security in VANET is obtained by restricting the unnecessary flow of information like vehicle id or position to other ongoing vehicles on the roadside. Tests also put forward that the penetration rate of the message does not affects its performance and reliability. As explained throughout the test section, this case is because of the way the traffic density is calculated, which directly relies on the number of slow vehicles detected.

In this paper, the following main further research topics have come up 1)the first topic will focus on modifying the vanet parallel ad dissemination architecture to use virtual segments or clustering techniques instead of the segments provided by a digital map to divide the road during the CEP processing. This way, the problems that might arise when the road segments are long could be sorted out. 2) FUZZY LOGIC based reputation system will be enhanced to punish the malicious nodes

8. REFERENCES

[1] Fernando Terroso, Rafael, "A Cooperative Approach To Traffic Congestion Detection With Complex Event Processing And Vanet,"in IEEE Transaction On Intelligent Transportation Systems, Vol 13,NO 2,JUNE 2012.

[2] Changqiao ,Hongke,"QoE-Driven User Centric VoD Services Multihomed P2P-Based Vehicular Networks",in IEEE Transactions on vehicular technology,vol 62,no 5,JUNE 2013

[3] Saif, Ali,"Context Aware Driver Behavior Detection System in Intelligent Transportation Systems (ITS)",in IEEE Transaction On Vehicular Technology,

[4]Ramin,"DTN Routing Protocols for VANETs:Issues and Approaches", in IJCSI ,Vol 8,Issue 6,No 1,November 2011

[5] E. Strom, H. Hartenstein, P. Santa, and W. Wiesbeck, "Vehicular communications: Ubiquitous networks for sustainable mobility," *Proc. IEEE*,vol. 98, no. 7, pp. 1111–1112, Jul. 2010.

[6] D. Luckham, *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Reading, MA: Addison- Wesley, 2002.

[7] O. Etzion, "Towards an event-driven architecture: An infrastructure for event processing position paper," in *nRules and Rule Markup Languages for the Semantic Web*. Berlin, Germany: Springer-Verlag, 2005, pp. 1–7.

[8] M. P. Gardner, "Highway traffic monitoring," Committee on Highway Traffic Monitoring, Washington, DC, Tech. Rep., 2000.

[9] H.-Y. Cheng and S.-H. Hsu, "Intelligent highway traffic surveillance with self-diagnosis abilities," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1462–1472, Dec. 2011.

[10] R. Wang, L. Zhang, R. Sun, J. Gong, and L. Cui, "Easitia: A pervasive traffic information acquisition system based on wireless sensor networks,"*IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 615–621, Jun. 2011.

[11] S. Vaqar and O. Basir, "Traffic pattern detection in a partially deployed vehicular ad hoc network of vehicles," *IEEE Trans. Wireless Commun.*,vol. 16, no. 6, pp. 40–46, Dec. 2009.

[12] R. Bauza, J. Gozalvez, and J. Sanchez-Soriano, "Road traffic congestion detection through cooperative vehicle-to-vehicle communications," in *Proc. 4th IEEE LCN Workshop User Mob. Veh. Netw.*, 2010, pp. 606–612.

[13] D. F. Llorca, M. A. Sotelo, S. Sánchez, M. Ocaña, J. M. Rodríguez- Ascariz, and M. A. García-Garrido, "Traffic data collection for floating car data enhancement in V2I networks," *EURASIP J. Adv. Signal Process.*,

vol. 2010, pp. 5:1–5:13, Mar. 2010.

[14] I. Leontiadis, G. Marfia, D. Mack, G. Pau, C. Mascolo, and M. Gerla, "On the effectiveness of an opportunistic traffic management system for

vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1537–1548, Dec. 2011.

[15] A. Skordylis and N. Trigoni, "Efficient data propagation in trafficmonitoring vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 680–694, Sep. 2011.

[16] M. Saito, J. Tsukamoto, T. Umedu, and T. Higashino, "Design and evaluation of intervehicle dissemination protocol for propagation of proceeding traffic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3,pp. 379–390, Sep. 2007.

[17] A. Adi, D. Botzer, G. Nechushtai, and G. Sharon, "Complex event processing for financial services," in *Proc. IEEE Serv. Comput. Workshops*, 2006, pp. 7–12.

[18] N. Museux, J. Mattioli, C. Laudy, and H. Soubaras, "Complex event processing approach for strategic intelligence," in *Proc. 9th Int. Conf. Inf. Fusion*, 2006, pp. 1–8.

[19] J. Dunkel, A. Fernández, R. Ortiz, and S. Ossowski, "Event-driven architecture for decision support in traffic management systems," *Expert Syst. App.*, vol. 38, no. 6, pp. 6530–6539, Jun. 2011.

[20] R.V.Aswiga, R.Praveena, S.jayanthi, R.Maheswari," Parallel Ad-Dissemination With Congestion Information And Driver Behavior Monitoring In Intelligent Vehicular Networks,"IJESIT,vol 2 ,issue 5 ,sept 2013.