International Journal of Computer Applications Technology and Research
Volume 2– Issue 3, 214 - 217, 2013, ISSN: 2319–8656

A HYBRID MODEL FOR MINING MULTI DIMENSIONAL DATA SETS

Santhosh kumar
Government College for Women (A),
PRIST University, Kumbakonam,
Tamil Nadu, India

E.Ramaraj
School of Computing
Alagappa University, Karaikudi,
Tamil Nadu, India

Abstract: This paper presents a hybrid data mining approach based on supervised learning and unsupervised learning to identify the closest data patterns in the data base. This technique enables to achieve the maximum accuracy rate with minimal complexity. The proposed algorithm is compared with traditional clustering and classification algorithm and it is also implemented with multidimensional datasets. The implementation results show better prediction accuracy and reliability.

Keywords: Classification, Clustering, C4.5, k-means Algorithm.

1. INTRODUCTION

Clustering and classification are the two familiar data mining techniques used for similar and dissimilar grouping of objects respectively. Although, due to efficient use of data mining techniques, clustering and classification techniques are used as pre-process activities. The clustering technique categories the data and reduces the number of features, removes irrelevant, redundant, or noisy data, and forms the sub groups of the given data based on its relativity. The classification is used as secondary process which further divides the similar (clustered) groups in to two discrete sub groups based on the attribute value. From our research work, especially for large data bases, the prior classification is required to minimize multiple features of data so that it can be mined easily. In this paper we proposed a combined approach of classification and clustering for gene sub type prediction.

2. PRELIMINARIES

2.1 C4.5 Algorithm

It is used to generate a decision tree developed by Ross Quinlan and it is an extension of ID3 algorithm. The decision trees generated by C4.5 can be used for classification, anC4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector, where they represent attributes or features of the sample, as well as the class in which falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sub lists.

2.2 K-means Algorithm

The k-means algorithm was developed by Mac Queen based on standard algorithm. It is one of the most widely used hard clustering techniques. This is an iterative method where the specified number of clusters should initialise earlier. One must specify the number of clusters beforehand. The algorithm can be specified as a given set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where μi is the mean of points in S_i
The algorithm works as follows:

- The k (number of clusters) value number of clusters must be initialised
- Randomly select k cluster centres (centroids) in the data space
- Assign data points to clusters based on the shortest Euclidean distance to the cluster centers
- Re-compute new cluster centers by averaging the observations assigned to a cluster
- Repeat above two steps until convergence criterion is satisfied

The advantage of this approach is its efficiency to handle large data sets and can work with compact clusters. The major limitation of this technique is the requirement to specify the number of clusters beforehand and its assumption that clusters are spherical.

3. RELATED STUDIES

In the year 2009 CSVM [1], Clustering based classification technique is proposed by Juanying Xie, Chunxia Wang, Yan Zhang, Shuai Jiang for unlabelled data prediction. They combined different kinds of k-means algorithm with SVM classifier to achieve better results. In order to avoid the major drawback of k-means algorithm; k-value initialisation, the CSVM is proposed. In 2010 [2], Pritha Mahata proposed a new hierarchical clustering technique called ECHC (exploratory consensus of hierarchical clustering's) which is used to sub group the various types of melanoma cancer. This work reveals that, k-means algorithm gives better results for biological subtype with proper sub tree. In 2010 [3], Taysir Hassan A. Soliman, proposed a clustering and classification as a combined approach to classify the different types of diseases based on gene selection method. The results had shown improved accuracy in data prediction. In 2011[4], Reuben Evans, Bernhard Pfahringer, Geoffrey Holmes, proposed a technique called statistical based clustering technique for large datasets. They used k-means algorithm as initial step for centroid prediction and classification as secondary step. In march 2013[5] Claudio Gentile, Fabio Vitale, Giovanni Zappella, implemented the combined technique (clustering and classification) in networks using signed graphs (classification) and correlation based grouping (clustering). In

this work we proposed that for very larger multi dimensional data base needs discrete classification as a primary process and its results classified samples that are grouped with the use of clustering methods. For classification the C4.5 classifier is used and k-means clustering is used as a secondary process.

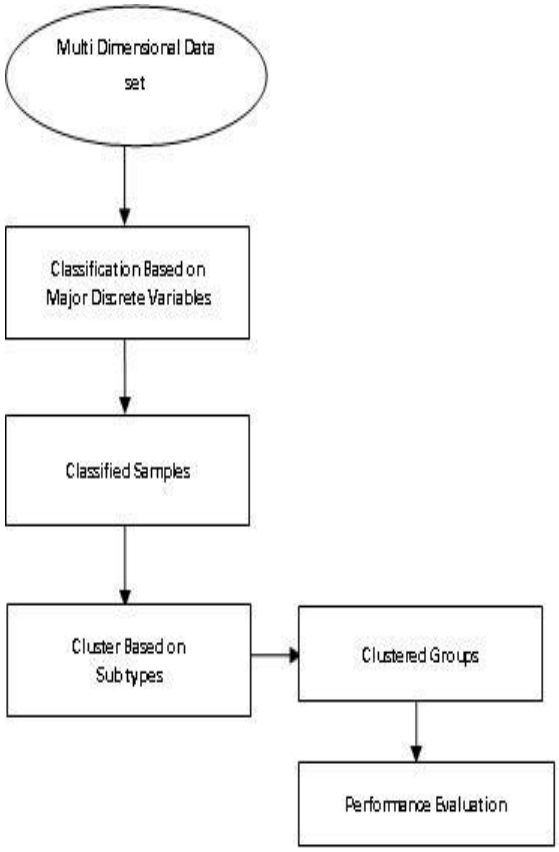

Fig 1: Hybrid model for Clustering and classification

4. PROPOSED WORK

With the combination of supervised and unsupervised learning approach, we proposed that

- A combined approach in needed in order to categorize and search the relevant data from multidimensional datasets
- For multi dimensional data sets, the primary partitioning is needed ; so that the categorized search gives better results with accuracy and time efficiency
- The classified samples can be sub grouped into smaller sets; so that the clustering can be done with global and local optimization

4.1 Proposed Model

Based on the drawbacks in existing approaches, we proposed a hybrid model for multi dimensional data sets. The major drawback of large datasets is its dimensionality. It is unavoidable, when needed information contains additional or unrelated data then it leads more difficult to search, analyze, and transfer the data. To overcome the above said problem, the major categorization of data based on requirement is needed. So that major part of large data can be divided in to two or more groups. One category contains the required data and other contains unrelated data respectively. We divided our work in to two parts; first approach is to divide the large data set in to major categories (based on its correlation) which requires discrete values to classify the data. For that a classifier should contain the data with related information in hierarchical form is needed. The C4.5 Decision tree algorithm is recommended to classify the data set into major discrete groups. Our second approach is to find the relativity among the categorized data. For that similar and dissimilar distance measures between the data items are considered. Based on its weight of the attributes in the group are organized. The k-means algorithm is used for clustering the similar data attributes in a categorized group.

5. EXPERIMENTAL EVALUATION

We have used colon cancer data set of colorectal cancer data from Biological data analysis web site [6]. The Data set contains 7,460 genes and 272 samples which contains both normal and infected genes. For our approach we have taken entire genes with samples and as stated above as a first approach we used C4.5 classifier to categorize data in terms of normal and infected genes based on its intensity range. The categorized genes are then clustered with the use of k-means algorithm. As a first step we implemented the data set into C4.5 classifier. The parameters taken for classification are mentioned below.

Table1. Classifier Parameter Initialization

Decision Tree (C 4.5)	**Top**
Min size of leaves	10
Confidence-level	0.25
Error Rate	0.8696

The following figure shows the graphical representation of major classification of genes as normal and tumor intensity.

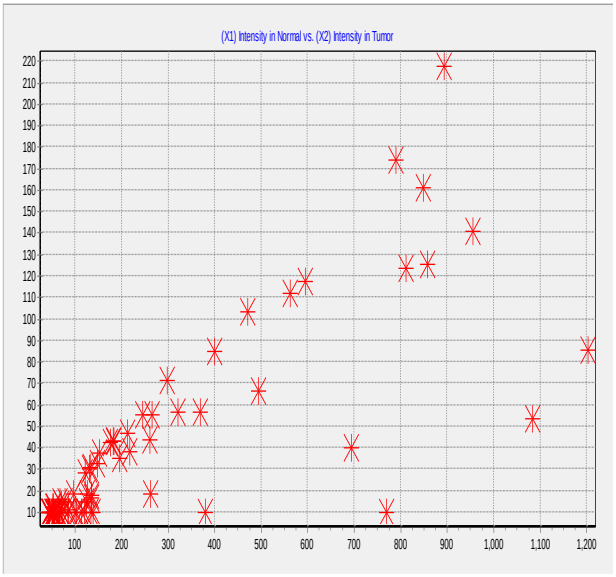

Fig.2. Normal Vs Tumor Intensity Classifier Representation

The second step is the given dataset is implemented directly into k-means algorithm for clustering. The parameters taken for clustering is tabled below.

Table.2. Cluster Parameter Initialization

Clusters (k-Means)	10
Max Iteration	10
Trials	5
Distance Normalization	variance

The figure shows the similar and dissimilar gene representation using k-means algorithm.

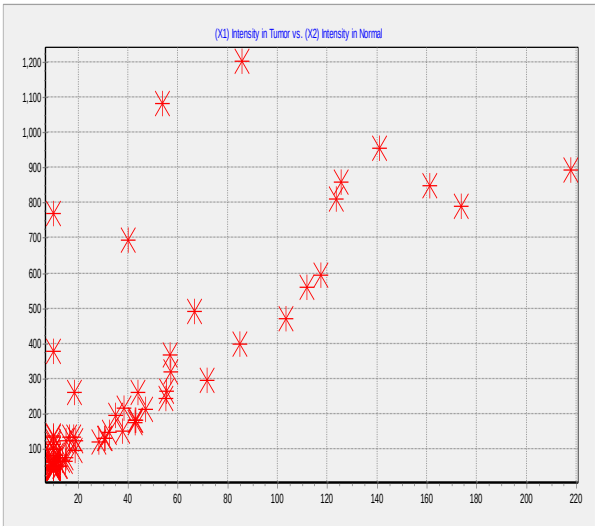

Fig.3. Normal Vs Tumor Intensity Cluster Representation

The final step of our work is the gene type is implemented using C4.5 classifier and the output obtained after classification is given as input for clustering process. The result is compared with individual implementation of previous steps. In order to evaluate the performance of hybrid method; the same parameters are applied for the proposed technique. The following graph shows the classification result of hybrid approach.

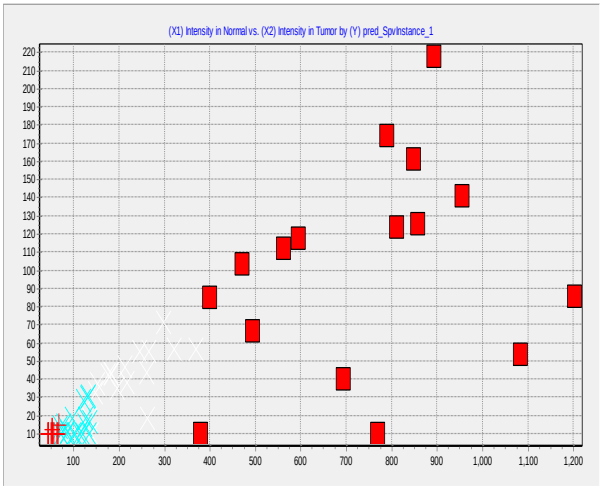

Fig.4. Normal Vs Tumor Intensity Hybrid Classifier Representation

The above graph denotes that the given genes are classified in to major categories which contain relevant attributes in each category.The following graph gives clear identification of similar and dissimilar representation of categorized groups. The irrelevant attributes are also considered and generated as separate clusters. Whereas the error rate based on the minimal size clusters are become comparatively less when it is compared with overall cluster centers.

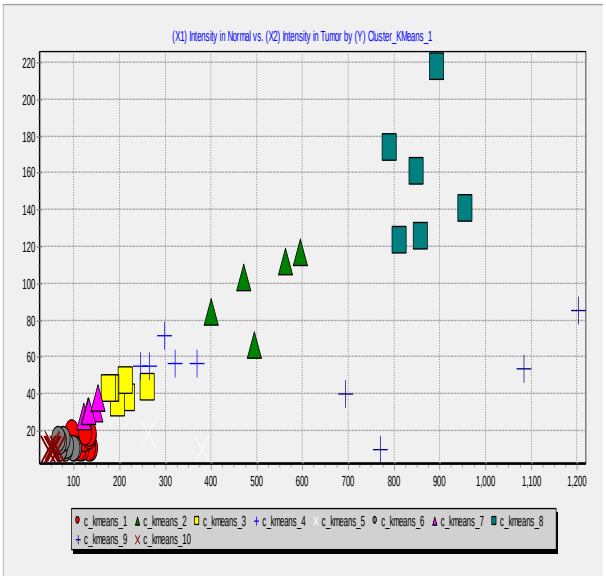

Fig.5. Normal Vs Tumor Intensity Hybrid (Classifier & Cluster) Representation

The experimented results error rate, computation time is analyzed and compared with the individual methods. The result shows that hybrid approach attains better results with less computation time. According to error rate, all the given attributes are represented as relevant or irrelevant groups so that the error rate must very less in hybrid approach.

Table3. Performance of Various Hybrid approach with Different groups methods

Method	Data Set	No. of Clusters/ D Tress	Computation Time
C 4.5	Colon	7 Nodes, 4 Leaves	31 Ms
K-Means	Colon	10 Clusters	15 Ms
Hybrid	Colon	10 Clusters	16 Ms

The table 3 shows that computation time of C4.5 classifier of using colon gene dataset is 31 milliseconds. The cluster using the same data set results 15 Milliseconds. But the combination of both C4.5 and k-means results 16 milliseconds. It reveals that the hybrid approach gives accurate results in less computation time.

6. CONCLUSION

From the analysis based on experimental results, we can conclude that our hybrid approach for multi dimensional data base gives better results in comparison with existing approaches. The large data sets are categorized using C4.5 classifier produces decision tree with relevant and irrelevant attributes in a set. This phenomenon makes possible to group the similar attributes in to group called cluster. Based on the hybrid approach one can divide the large database in to major groups. The major groups can be further clustered in to similar group which enables to achieve high accuracy rate with less computation time. This hybrid approach is suitable for large data bases having multi dimensional complexity. By using this approach one can retrieve exact information from

the data base with in stipulated time by removing or minimizing the additional features of large data.

7. REFERENCES

[1] **Juanying Xie** ; **Chunxia** Wang ; Yan **Zhang,** 2009, ICTM **2009** conference.

[2] Pritha Mahata, January-march 2010. Ieee/acm transactions on computational biology and bioinformatics, vol. 7, no. 1,

[3] Taysir Hassan A. Soliman, Adel A. Sewissy, and Hisham Abdel LatifSannella, 2010, A gene selection approach for classifying diseases based on microarray datasets , IEEE 12th International Conference on Bioinformatics and Bioengineering, Cyprus

[4] Reuben Evans, Bernhard Pfahringer, Geoffrey Holmes. 2011, IEEE, Clustering for classification.

[5] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[6] http://www.ncbi.nlm.nih.gov/gene

[7] Chenn-Jung Huang ,Wei-Chen Liao, "A Comparative Study of Feature Selection Methods for Probabilistic Neural Networks in Cancer Classification", Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03),Vol 3, pp1082-3409, 2003.

[8] http://sdmc.lit.org.sg/GEDatasets/

[9] Debahuti Mishra, Barnali Sahu, 2011,"A signal to noise classification model for identification of differentially expressed genes from gene expression data,"3rd International conference on electronics computer technology.

[10] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," SIGKDD Explor, Vol. 6, 2004, pp. 90-105.

[11] G. Moise, A. Zimek, P. Kroger, H.P. Kriegel, and J. Sander, "Subspace and projected clustering: experimental evaluation and analysis," Knowl. Inf. Syst., Vol. 3, 2009, pp. 299-326.

[12] H.P. Kriegel, P. Kroger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," ACM Trans. Knowl. Discov. Data., Vol 3, 2009, pp. 1-58.

[13] ErendiraRendon, Itzel Abundez, Alejandra Arizmendi, Elvia M. Quiroz, "Internal versus External cluster validation indexes," International Journal of Computers and Communications, Vol. 5, No. 1, 2011, pp. 27-34.

[14] Bolshakova, N., Azuaje, F., "Machaon CVE: cluster validation for gene expression data,"Bioinformatics, Vol. 19, No. 18, 2003, pp. 2494-2495.