Impact of Bio-inspired metaheuristics in the data clustering problem

K. Sumangala,
Assistant Professor, Research Scholar,
Department of Computer Science,
Kongunadu Arts & Science College,
Coimbatore - 641 030, India
.

K. Papitha
Department of Computer Science,
Prince Shri Venkateswara Arts & Science
College,Chennai
India

Abstract: The goal of data mining is to extract the knowledge from data. It is also a form of knowledge discovery essential for solving problem in a specific domain. This paper presents a novel approach to data clustering and classification problem. Clustering analysis is distribution of data into groups of similar objects and Classification focuses the data on the class boundaries. This research explores three different bio-inspired metaheuristic algorithms in the clustering problem: Ant Colony Optimization (ACO), Genetic Algorithms (GAs) and Artificial Immune Systems (AIS). Data mining approaches are applied in the field of medical diagnosis recently. The major class of problem in medical science involves diagnosis of disease based upon various tests. The computerized diagnostic tools are helpful to predict the diagnosis accurately. Breast cancer is one the most dangerous cancer type in the world. Early detection can save a life and increase survivability of the patients. This of research work analysed the performance of GA, ACO and AIS with ID3 for solving data clustering and classification problem in an experiment with Breast Cancer Dataset data of UCI repository. An efficient ID3 Decision tree based classification techniques are used to measure the performance of the system with GA, ACO and AIS system. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers. Instead of K-means clustering, this research work combines the simplicity of K-means algorithm with the robustness of AGA-Miner. This proposed approach has potential applications in hospital for decision-making and analyze/ research such as predictive medicine.

Keywords: Clustering problem; genetic algorithms; ant colony optimization; artificial immune systems, ID3 classification techniques, UCI data repository.

1. INTRODUCTION

The goal of data mining is to extract the knowledge from data. It is also a form of knowledge discovery essential for solving problem in a specific domain. This paper presents a novel approach to data clustering and classification problem. Clustering analysis is distribution of data into groups of similar objects and Classification focuses the data on the class boundaries. This research explores three different bio-inspired metaheuristic algorithms in the clustering problem: Ant Colony Optimization (ACO), Genetic Algorithms (GAs) and Artificial Immune Systems (AIS). Data mining approaches are applied in the field of medical diagnosis recently. The major class of problem in medical science involves diagnosis of disease based upon various tests. The computerized diagnostic tools are helpful to predict the diagnosis accurately. Breast cancer is one the most dangerous cancer type in the world. Early detection can save a life and increase survivability of the patients. This of research work analysed the performance of GA, ACO and AIS with ID3 for solving data clustering and classification problem in an experiment with Breast Cancer Dataset data of UCI repository. Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms. It uses information gain to select best attribute for splitting. An efficient ID3 Decision tree based classification techniques are used to measure the performance of the system with GA, ACO and AIS system. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers. Instead of K-means clustering, this research work combines the simplicity of K-means algorithm with the robustness of AGA-Miner. This proposed approach has potential applications in hospital for decision-making and analyze/ research such as predictive medicine. In this study we have developed AGA-Miner that selects the best cluster centroid value than the existing clustering methods.

The rest of this paper is organized as follows. Section 2 presents the data clustering problem, standard K-means algorithm and related Data mining techniques. Section 3 addresses the bio-inspired metaheuristics: ACO, GA, AIS and ID3. Section 4 explains empirical studies performed using these metaheuristics on data clustering and Section 5 compares the different approaches of AGA-Miner. Finally, concluding remarks are given in Section 6.

2. CLUSTERING ALGORITHM

The process of grouping a set of abstract objects into classes of similar objects called clustering [1]. The clustering problem can be described as follows:

$$J(W,C) = \sum_i^N \sum_j^K w_{ij} \left|\left|x_i - c_j\right|\right|^2 \quad (2.1)$$

$$\sum_j^k w_{ij} = 1 \quad (2.2)$$

$$C_j = \frac{1}{N_j} \sum_{\pi \in Cj} x_i \quad (2.3)$$

Where, K-> number of cluster, n-> number of objects , m-> number of attribute, Cj->center of j^{th}cluster, x_i->location of i^{th} object.

Euclidean distance:
The distance between the data vector and centroid C is calculated by:

$$\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \qquad (2.4)$$

2.1. K-Means Algorithm:

One of the most widely used algorithms is k-means clustering. It partitions the objects into clusters by minimizing the sum of the squared distances between the objects and the centroid of the clusters. The k-means clustering is simple but it has high time complexity, so it is not suitable for large data set. One of the most popular clustering techniques is the k-means clustering algorithm.

Starting from a random partitioning, the algorithm repeatedly (i) computes the current cluster centers (i.e. the average vector of each cluster in data space) and (ii) reassigns each data item to the cluster whose center is closest to it.

The algorithm for the standard k- means clustering is given as follows [2]:

a. Choose a number of clusters k
b. Initialize cluster centers µ1,… µk
  i. Could pick k data points and set cluster centers to these points
  ii. Or could randomly assign points to clusters and take means of clusters.
c. For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster.
d. Re-compute cluster centers (mean of data points in cluster)
e. Stop when there are no new re-assignments.

Advantage:

- Simple and widely used clustering algorithm.

Disadvantage:

- Time complexity (when large dataset are to be clustered)
- Need to estimate the number of cluster in advance.

2.2. Data Mining Techniques

This research work presents a novel approach to data clustering and classification problem [2]. The clustering is to discover the data distribution. The proposed system is able to cluster real value data efficiently and correctly, dynamically estimating number of cluster. This research work analyzed the performance of ACO, GA and AIS metaheuristics algorithms. In classification problem discrimination among classes is based on the decision tree ID3 classifier.

- Decision Tree (ID3)
- Genetic Algorithm (GA)
- Ant Colony Optimization (ACO)
- Artificial Immune System (AIS)

Decision Tree (ID3):

- Iterative Dichotomiser 3 (ID3) is algorithm for building decision tree.
- It uses information gain to select best attribute for splitting.

Genetic Algorithm (GA):

- **GA** is a technique for solving the clustering problem.
- **GKA** combines K-Means with GA to find globally partition for dataset into number of cluster.
- It uses process such as population, crossover, mutation.
- ID3 algorithm applied to classify the cancer dataset result from the cluster result. ID3 selects the test attribute based on Information Gain.

Ant Colony Optimization (ACO):

- It combines K-means with ACO to improve the k-means in two steps:
  - To avoid local optima
  - ACO applied to refine the cluster to improve quality.
- Ants are used to cluster the data points.
- Only one ant is used to refine the cluster.
- Whenever it crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving with the help of pickup and drop up probabilities.

Artificial Immune System (AIS):

- AIS is try to imitate real immune system. Most AIS use only the main ideas of real immune systems, namely clonal and negative selection which deal with the evolution of B-cells and T-cells, respectively.
- The proposed method belongs to the method derived from immune system paradigm, called ClonalG Selection.
- It resembles the original K-Means algorithm, but it get rid of its main drawback
  - It is able to estimate the proper number of cluster & avoids getting stuck in inappropriate areas.

- Comparing to other immune algorithms for data clustering, its computational cost is decreased by producing a limited number of clones and proper suppression mechanism.

3. METAHEURISTICS GA, ACO, AIS AND ID3 BASED METHODS

Heuristics refers to experience-based techniques for problem solving, learning, and discovery [1]. In computer science, metaheuristicis a computational method that optimizes a problem by iteratively trying to improve a candidate solution. Metaheuristics allows us to find the best solution over a discrete search-space.

3.1 Classification Techniques (ID3):

The decision tree is constructed with each non-terminal node representing the selected attribute on which data was split and terminal nodes representing the class label of the final subset of its branch. ID3 is an algorithm for building decision tree. It uses information gain to select best attribute for splitting. Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms [17]. The main purpose of decision tree is to expose the structural information contained in the data. If the target variable (also called as response variable or class variable) is nominal/categorical variable is called “classification tree” and if continuous, the tree is called “regression tree”. ID3 is a recursive process used to construct decision tree from data.

Building decision tree:

1. Calculate the entropy for every attribute using the data set S.
2. Split the set S into subsets using the attribute for which entropy is minimum. Make a decision tree node containing that attribute.
3. Recurse on subsets using remaining attributes. Stop splitting when all examples at a node have the same labels.

Entropy:

- Entropy is a quantitative measurement of the homogeneity of a set of examples.
- It tells us how well an attribute separate the training examples according to their target classification.
- Given a set S with only two class case (malignant & benign)

$$\text{Entropy}(S) = -P_m \log_2 P_m - P_b \log_2 P_b \qquad \textbf{(3.1.1)}$$

Where P_m = proportion of malignant examples
P_b = proportion of benign examples

If entropy(S) = 0, all members in S belongs to one class.
If entropy(S) = 1(max value), members are Split equally between two classes.

In general, if an attribute takes more than two values

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log(p_i) \qquad (3.1.2)$$

P**seudo code of ID3 :**

```
ID3 ( Learning Sets S, Attributes Sets A, Attributes values V)

Begin
        Load learning sets first, create decision tree root node
'rootNode', add learning set S into root node as its subset.
For root Node, we compute Entropy (rootNode.subset) first

        If Entropy(rootNode.subset)==0, then
                rootNode.subset consists of records all with
                the same value for the categorical attribute,
                return a leaf node with decisionattribute:
                attribute value;

        If Entropy(rootNode.subset)!=0, then
                compute information gain for each attribute
                left(have not been used in splitting), find
                attribute A with Maximum(Gain(S,A)).

                Create child nodes of this rootNode and add
                to rootNode in the decision tree.

        For each child of the rootNode, applyID3(S,A,V)
        recursively until reach node that has entropy=0 or reach
        leaf node.
End ID3.
```

- Looking for which attribute creates the most homogeneous branches

$$\mathbf{Gain}(\boldsymbol{S},\boldsymbol{A}) = \boldsymbol{Entropy}(\boldsymbol{S}) - \sum_{v \in \boldsymbol{Value}(\boldsymbol{A})} \left|\frac{\boldsymbol{S_v}}{\boldsymbol{S}}\right| \boldsymbol{Entropy}(\boldsymbol{S_v}) \qquad (3.1.3)$$

Where, A is an attribute of S, Value(A) is the set of possible value of A, v is a particular value in Value(A), S_v is a subset of S having of v’s on value(A)

3.2. GA with clustering & classification:

- **GA** is a technique for solving the clustering problem [8].
- **GKA** combines K-Means with GA to find globally partition for dataset into number of cluster [1]. The purpose of GKA (Genetic K-Means Algorithm) is to minimize intra-cluster variance.
- From the initial population, the basic operators (selection, crossover, mutation) evolve the population generation to generation. These operators has been used to populate the data

points which helps to find the best fitness solution.

- **Operations:**
  - **Population:** Initially, populate the Breast Cancer dataset attribute value as 0's and 1's.
  - **Selection: "Select the best, discard the rest".** Select the present attribute data & compute the fitness solution. The selection operations, selects the best fitness solution and stores into global.
  - **Crossover & Mutation:**
    - **Crossover:** Cross over operator combines parts of two part solutions to create new solution.
    - **Mutation:** Mutation operator modifies randomly the solution created by crossover.

Euclidean distance:

The distance between the data vector x and centroid c is calculated by,

$$\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \qquad (3.2.1)$$

Pseudo code of GA-Clustering Algorithm

```
Begin
    1.  t=0
    2.  Initialize population P(t)
    3.  Compute fitness P(t)
    4.  t=t+1
    5.  If termination criterion is achieved go to step 10
    6.  Select P(t) from P(t-1)
    7.  Crossover P(t)
    8.  Mutate P(t)
    9.  Go to step 3
    10. Output best and stop

End
```

- ID3 algorithm applied to classify the cancer dataset result from the cluster result.
- ID3 selects the test attribute based on Information gain. IG measures the change of uncertainty level after classification from an attribute.

3.3 ACO with clustering & classification:

The proposed system combines K-Means with Ant Colony Optimization to improve K-Means in two steps [7]:

1. To avoid local optima.
2. ACO applied to refine the cluster to improve quality.

Pseudo code of ACO Algorithm

```
ACO Algorithm:

    1. Choose number of cluster k
    2. Initialize cluster center $\mu_1, \mu_2 ..... \mu_n$.
    3. For each data points , compute the cluster center ,it is
       closet to and assign the data point to this cluster.
    4. Re-compute cluster center.
    5. Stop when no new reassignments.
    6. Ant based refinement:
        1. Input the cluster from improved K-means.
        2. For i=1 to N do
            1. Let the ant go for random walk to pick an item.
            2. Calculate pick up & drop up probability.
            3. Decide to drop the item.
            4. Re-calculate the entropy value to check whether the
               quality is improving.
        3. Repeat.
```

The ants are used to cluster the data points. Here, only one ant is used to refine the cluster. Whenever it crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving with the help of pick up and drop up probabilities.

Entropy:

The quality of cluster analyzed using two measures: Entropy, F-Measure. For each cluster, the class distribution of the data is calculated first.

$$E_j = -\sum_i p_{ij} \log(p_{ij}) \qquad (3.3.1)$$

Whenever the ant crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving.

Pickup probability $Pp = \left(\frac{k1}{k1+f}\right)^2$ (3.3.2)

Drop up probability $Pd = \left(\frac{f}{k2+f}\right)^2$ (3.3.3)

where f -> Entropy value, (Calculated before the item pick up & drop up)
k1 , k2 -> Threshold Constants.

- If Pd<Pp , item is dropped into another cluster and entropy value calculated again.

3.4. AIS with clustering & classification:

AIS is try to imitate real immune system. Most AIS use only the main ideas of real immune systems, namely clonal and negative selections which deal with the evolution of B-cells and T-cells, respectively [3]. The proposed method belongs to the method derived from immune system paradigm, called Clonal-G Selection.

- It resembles the original K-Means algorithm, but it get rid of its main drawback
    - It is able to estimate the proper number of cluster & avoids getting stuck in inappropriate areas.
    - Comparing to other immune algorithms for data clustering, its computational cost is decreased by producing a limited number of clones and proper suppression mechanism.
- Clonal Selection is used in data compression, data and web mining, clustering and optimization. Traditional clustering algorithm is to find the data distribution using cluster centers. When used for classification, these cluster centers are sometimes not the best ones, especially when the number of cluster is too small. The Clonal selection aims to evade this obstacle by means of a new suppression mechanism, which focuses learning on the boundaries among classes.
- **Concept:**
    - B-cells with different receptors' shapes try to bind to antigens (Training & Testing data).
    - The best fitted B-cells become stimulated and start to proliferate and produce clones, which are then mutated at very high rates.
    - After this process is repeated, it will emerge better B-Cells (Best Solution).

Pseudo code of AIS Algorithm:

1. A set of antigens Ag is presented to the antibodies population Ab ;
2. The affinity measure f of the antibodies in relation to the antigens is calculated;
3. The n highest affinity antibodies to the antigens are selected to be cloned,generating the antibody subset Ab {n} ;
4. The antibodies selected will be cloned according to their affinity to the antigens (as higher the affinity more clones it will generate), producing a C clones population;

5. The C clones population is subjected to an affinity maturation process at an inversely proportional rate to the affinity of the clone (as higher the affinity, lower the mutation rate), and a new population of clones C* is produced;
6. The C* clones population is evaluated and its affinity measure f* in relation to the antigens is calculated;
7. The n matured antibodies of the highest affinity are selected to compose the next population generation, since its affinity is greater than its original antibodies;
8. The d worst antibodies are removed from the population and replaced by new randomly generated antibodies.

This process repeats until a stop condition (number of generations) is reached.

- The population of B-cells depends on several mechanisms
    - Recognition
    - Stimulation
    - Proliferation
    - Hyper mutation
    - Suppression

Recognition:

Recognition of antigens (Training & Testing data) by B-cells depends on level of binding between them.

Stimulation:

The level of binding can be stimulated by given metric (Eg: Euclidean distance).

Hyper mutation:

It can be easily done by random changes in feature vector describing B-cells.

Suppression:

It is playing a vital role in the processing.

Clonal-G Selection:

- Simple
- Idea:
    - Clone generation (Choose number of attributes)
    - Suppression mechanism (Remove useless cluster's center).

4. RESULT AND DISCUSSION

4.1 GAC and MAC Empirical studies

GAC was executed with varied parameters values on the Breast database in order to choose the best setting among the existing possibilities, such as, mutation operator, crossover operator, mutation and crossover rates [1].

In these experiments the following data were analyzed: (a) Best solution; (b) Worst solution; (c) Average of the best solutions; (d) Standard deviation; (e) Average number of objective function evaluations.

Table 4.1 K-Means with Genetic

Database	Algorithm	Fitness of the Best Solution	Fitness of the Worst Solution	Average Fitness of the Solutions	Standard Deviation	Average of the Fitness Evaluations
Breast	ACO	334.00	334.00	335.00	0.90	23,060.00

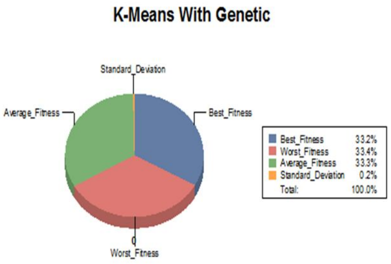


Figure-4.1: Graph result of GAC

4.2 ACO Empirical Studies

- Ants are used to cluster the data points using entropy [1].
- Entropy compute the probability (Pi,j) that the member of cluster belongs to the class.

Table 4.2 K-Means with ACO

Database	Algorithm	Fitness of the Best Solution	Fitness of the Worst Solution	Average Fitness of the Solutions	Standard Deviation	Average of the Fitness Evaluations
Breast	ACO	334.00	334.00	335.00	0.90	23,060.00

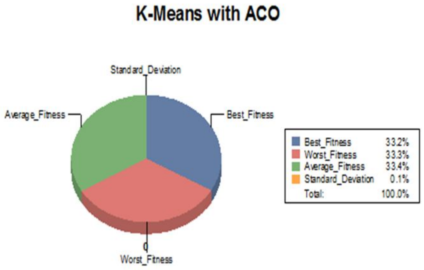


Figure-4.2: Graph result of ACO

4.3 AIS Empirical Studies

- The algorithms were executed with varied input parameter values on the breast database [1].
- In CLONLG algorithm, number of cluster, number of attributes to choose will be defined.

Database	Algorithm	Fitness of the Best Solution	Fitness of the Worst Solution	Average Fitness of the Solutions	Standard Deviation	Average of the Fitness Evaluations
Breast	AIS	383.00	1,194.00	764.00	18.5	23,189.00

Table: 4.3 K-Means with AIS

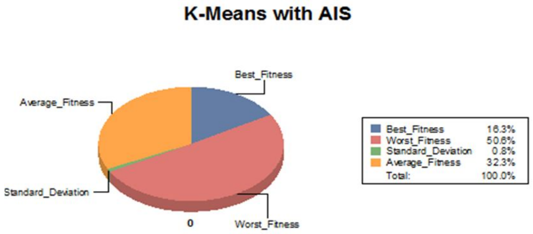


Figure-4.3: Graph result of AIS

4.4 ID3 Empirical Studies

- ID3 is a recursive process used to construct decision tree from data.
- Classification Trees are methodologies to classify data into discrete ones using the tree-structured algorithms. The main purpose of decision tree is to expose the structural information contained in the data.
- If the target variable (also called as response variable or class variable) is nominal/categorical variable is called "classification tree" and if continuous, the tree is called "regression tree" .

Building decision tree:

- Calculate the entropy for every attribute using the data set S.
- Split the set S into subsets using the attribute for which entropy is minimum. Make a decision tree node containing that attribute.
- Recurse on subsets using remaining attributes. Stop splitting when all examples at a node have the same labels.

- Entropy
  - used to measure the uncertainty associated with a random variable
  - Entropy(S) = $\sum_{i=1}^{n} -p_i \log(p_i)$
- Information gain
  - Information gain is based on the decrease in entropy after a dataset is split on an attribute.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \left|\frac{S_v}{S}\right| Entropy(S_v) \quad (4.4.1)$$

The performance of each algorithm is measured with the best, worst and normal fitness values count of each algorithm with the Euclidean distance measure for searching process. The Breast cancer dataset measure the standard deviation and mean average distance value of algorithm finally proposed AIS with ID3 shows the best classification result than the existing GA with ID3 and ACO with ID3 algorithm in breast cancer classification with fitness values.

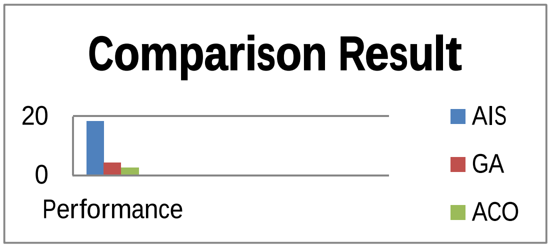
5. COMPARISON RESULT

In this section we measure the performance of the system and show the results of the accuracy in terms of how clustering performance is improved. The Breast cancer dataset measure the standard deviation and mean average distance value of algorithm finally proposed AIS with ID3 shows the best classification result than the existing GA with ID3 and ACO with ID3 algorithm in breast cancer classification with fitness values.

Table 5.1 Comparison result of ACO,GA,AIS

Database	Algorithm	Fitness of the Best Solution	Standard Deviation	Average of the Fitness Evaluations
Breast	GA	337.00	1.70	23,330.00
	ACO	334.00	0.90	23,060.00
	AIS	383.00	18.5	23,189.00

Figure 5.1 Comparison result of ACO,GA,AIS



6. CONCLUSION

In this work analyzed the performance of the GA, ACO and AIS metaheuristic algorithms with ID3 for solving data Clustering & Classification problem in an experiment with breast cancer dataset data of UCI repository. Instead of K-Mean clustering, this research work can combine the simplicity of the K-Means algorithm with the robustness of AGA-Miner.

In the proposed system, an efficient ID3 classification techniques are used to measure the performance of the system with GA, ACO and AIS system, which could increase the accuracy and reduces the cost of time. Proposed AIS system produces the best classification result than the ACO and GA based decision tree ID3 classifiers.

6.1 FUTURE ENHANCEMENT

As per our observation there are some future suggestions which are listed below:

- This experimental result may be used to detect other cancer types such as lung cancer, mouth caner and etc.
- Apply the other classification algorithm such as C4.5 and C5.0, Ripper classification algorithm and compare the results with ID3 methods.

Instead of applying the optimization methods different correlation based similarity measures with optimization are applied to cluster the dataset.

7. REFERENCES

[1] Ana Cristina B.Kochem Vendramin, Diogo Augusto Barros Pereira, “Application of Bio-inspired Metaheuristics in the data clustering problem”

[2] Benny Pinkas, Yehuda Lindell, “Privacy Preserving Data Mining”.

[3] Berkhin, P. 2002. “Survey clustering data mining techniques”, Technical report, Accrue software, San Jose, California.

[4] M. Bramer. “Principles of Data Mining”.Springer, 2007.

[5] D. Dasgupta, Artificial Immune Systems and Their Applications, Springer, Berlin, 1999 .

[6] Ding, C., and He, X. 2002. “Cluster merging and splitting in hierarchical clustering algorithms”, IEEE international conference, pp. 139-146.

[7] Dorigo, M., Maniezzo.V,& Colorni .A.,” Ant System: Optimization by a colony of cooperating agents,” IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26, 29–41,1996.

[8] G. Garai and B. B. Chaudhuri. “A Novel Genetic Algorithm for Automatic Clustering”. Pattern Recognition Letters, vol. 25, n.2, pp.173-187, 2004.