International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012

DOI: 10.7321/jscse.v2.n5.2

An Ant Colony Clustering Algorithm Using Fuzzy Logic

[1]S.Nithya, [2]R.Manavalan
[1] *M.Phil Scholar, Full Time, Dept of Computer Science, K.S.R College of Arts and Science, Tiruchengode*
[*] *Department of Computer Science and Applications, K.S.R College of Arts and Science, Tiruchengode,*
[1]*nithu_nivi@yahoo.co.in,* [2] *manavalan_r@rediffmail.in*

Abstract. **The performance of Data partitioning using machine learning techniques is calculated only with distance measures i.e similarity between the transactions is carried out with the help of distance measurement algorithms such as Euclidian distance measure and cosine distance measure. The distance with connectivity (DWC) model is used to estimate distance between transactions with local consistency and global connectivity information. The ant colony optimization (ACO) techniques are used for the data clustering process. In this paper we propose distance measure model of DWC by enhancing the model using fuzzy logic. The transaction weights are updated using fuzzification process. All the attribute weight values are updated with a fuzzy set weight value. The distance with connectivity model is tuned to estimate distance between the transactions using the fuzzy set values. The distance measure model efficiently handles the uneven transaction distributions. The ant colony-clustering algorithm is also improved with fuzzy logic. The similarity computations are carried out with fuzzy distance measurement models. Un-even data distribution handling, accurate distance measure and cluster accuracy are the features of the proposed clustering algorithm.**

Keywords: *Distance with connectivity, Ant colony optimization, Fuzzification, Fuzzy Ant Colony Optimization, Breast Cancer Dataset.*

1. Introduction:

Clustering plays a significant role in the field of research. Cluster is a collection of objects which are “similar” among them and are “dissimilar” to the objects belonging to other clusters. To help us better in our path of knowing more about clustering none other than Ants play an important role for identifying the clusters [1].

Ant colony optimization (ACO) is a population-based metaheuristic that can be used to find approximate solutions to difficult optimization problems. The ant colony clustering algorithm imitates the intelligent behavior of ant and applies it to the solution of hard computational problems. This work was initiated by Deneubourg [2]. To apply ACO, the ant randomly moving and picking, dropping the object to achieve

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

cluster. [3], [4], [5], [6]. The optimization problem is finding the best path on a weighted graph. The artificial ants incrementally build solutions by moving on the graph. The solution construction process is stochastic and is biased by a pheromone model, that is, a set of parameters associated with graph components (either nodes or edges) whose values are modified at runtime by the ants.

In traditional ways like Euclidean distance and cosine distance were used to find out the connectivity among data, but the limitation involved was it can be done only in local consistency and not in global connectivity. Therefore the appropriate solution was to go for DWC a distance calculating method [7]. One of the merits of DWC is to find inherent clustering characters of the data sets and suitable for data sets with uneven density distribution, comparing with the density-based methods. The overview of proposed work as follows:

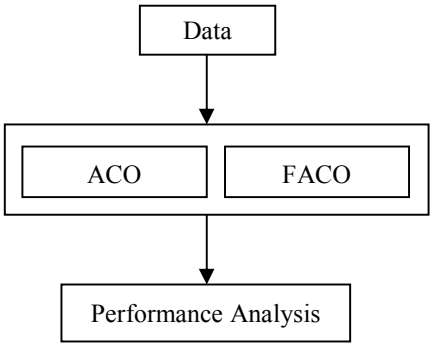

Figure1. Proposed Model

This study is needed to overcome the local consistency problem and to find the distance in global connectivity and further the accuracy has been increased by coalescing fuzzy logic with ACODWC. This paper emphasis on clustering model to increase the accuracy and solving the local optimum problem.

This paper is organized as follows: section 2 describes Literature survey, section 3 describes regarding Data Connectivity based distance estimation, section 4 depicts about Ant colony clustering algorithm based on DWC, Solution construction and Pheromone update rule, section 5 portrays on Fuzzy enabled clustering scheme, fuzzy logic concepts, Distance analysis, Fuzzification process, Fuzzy ACO, section 6 express in relation to Experimental analysis, section 7 illustrate with reference to Evaluation method, section 7 is end with wrapping up of the current work and upcoming enrichment.

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

2. Literature Survey

Fuzzy ants and clustering projected a method Fuzzy c-means and hard c-means for reformulated the fuzzy partition validity metric and to clustering the data. Clustering web search results using fuzzy ants anticipated a method ACO simulated by means of fuzzy IF-THEN rules or fuzzy logic for clustering web search results have to be efficient and robust and cluster a dataset without using any kind of a priori information. Fuzzy controller design by clustering-aided ant colony optimization proposed a method Ant colony optimization (ACO) algorithm (CACO) for improving both the design efficiency of a fuzzy controllers and its performance. Fuzzy ant based clustering proposed ACO and fuzzy if–then rules for clustering the data with no initial partitioning and the number of clusters to be known in initial. An effective clustering algorithm with ant colony a method namely Sacc algorithm and Jaccard index for solving unsupervised clustering problem and to identify the optimal cluster number. A fuzzy-ACO method called Fuzzy rules and ACO for Detecting breast cancer for solving problems in specific domain

3. Data Connectivity Based Distance Estimation

DWC is used to find the distance in local compactness and global connectivity between the data. Establish the adjacency matrix of the data set. First, calculate the set $\phi_i = \{\phi_{ij} \in X, j = 1,...L\}$ of the data object $x_i(i = 1,...N)$ L (L> 0) nearest-neighbor [8], using the Euclidean distance formula. Then, link xi and (i= 1,…,N) L (L> 0) , and the link is undirected. First should calculate nearest neighbor using the Euclidean distance formula [8]. In this way, we construct an undirected graph G = (X, V) and the adjacency matrix $R = [R_y]_{N \times N}$ of the data set, where:

$$R_{ij} = \begin{cases} 1, x_j \in \phi_i or x_i \in \phi_i \\ 0, otherwise \end{cases}$$

V is the set of all links between data. Rs = [Rij s]N × N , Rij s is the number of s-steps reachability paths between xi and xj .

Definition 1: The connectivity between data object xi and xj is defined as:

$$Conn(x_i, x_j) = \sum_{s-1}^{step} conn^s(x_i, x_j)$$

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

$$Conn^{s}(x_i, x_j) = \begin{cases} \dfrac{\log_L R_{ij}^{\,s}}{s-1}, if\ s > 1\ \wedge R_{ij}^{\,s} > 1 \wedge i \neq j \\ 1, if\ s = 1 \wedge R_{ij}^{\,s} > 0\ \ i \neq j \\ 0\ \ , otherwise \end{cases}$$

The higher the connectivity between data point xi and xj is, the more reachability paths between xi and xj will have, which reflects the higher similarity between xi and xj , and the more close distance of xi and xj . Hence, DWC (xi, xj) $\propto$ 1/Conn (xi, xj). Furthermore, the defintion of DWC (xi, xj) should reflect the local consistency at the same time, then we difined the DWC distance of data objects as follows.

Definition 2: The DWC distance between data object xi and xj is difined as:

$$Conn^{s}(x_i, x_j) = \begin{cases} \dfrac{\log_L R_{ij}^{\,s}}{s-1}, if\ s > 1\ \wedge R_{ij}^{\,s} > 1 \wedge i \neq j \\ 1, if\ s = 1 \wedge R_{ij}^{\,s} > 0\ \ i \neq j \\ 0\ \ , otherwise \end{cases}$$

Where

$$Dis(x_i, x_j) = \sqrt{\sum_{v=1}^{m} \left| x_{iv} - x_{jv} \right|^2}$$

It is the Euclidean distance between xi and xj, m denotes the number of the data object attributes, Max is a very large positive constant, M is a positive constant.

If $Dis\ (x_i, x_j)$ is short and $Conn\ (x_i, x_j)$ is high, $DWC\ (x_i, x_j)$ will be small. Then, the data object xi and xj will be clustered into the same cluster with a high probability. If $Dis\ (x_i, x_j)$ is short, but $Conn\ (x_i, x_j)$ is low, $DWC\ (x_i, x_j)$ will still be large. Then, object xi and xj will not be grouped in the same cluster.

Obviously, DWC satisfies the following basic properties:
• DWC (x, y) ≥ 0 , if and only if x = y ,equality holds;
• DWC (x, y) = DWC (y, x).
DWC does not always satisfy the triangle inequality, so the definition of DWC is a generalized distance.

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

4. Ant Colony Clustering Algorithm Based On DWC

The ant colony-clustering algorithm improved with DWC [3]. Given {x1,x2,….xN} a data set of N objects, and K(0<K<N), the number of clusters to form, clustering analysis organizes the N objects into K clusters, in order to minimize the clustering objective function F, where each object x1(i=1,….N) has m attributes, expressed as {xi1,xi2,….xim} [9].

The objective function is computed as follows:

$$\text{Min } F(w,C) = \sum_{j=1}^{K}\sum_{i=1}^{N} w_{ij} DWC(x_i, C_j) \quad (4)$$

Cj → Centroid of Clusters(j=1,…..,k), xi → Object

Subject to:

$$\sum_{j=1}^{K} w_{ij} = 1, i = 1,2,...., N - \quad (5)$$

$$\sum_{i=1}^{N} w_{ij} \geq 1, j = 1,2,...., K - \quad (6)$$

Here, w is an N-by-K weighting matrix, its elements:

$$w_{ij} = \begin{cases} 1, if x_i \in cluster_j \\ 0, if x_i \notin cluster_j \end{cases} \quad (7)$$

4.1. Solution construction

In ant colony algorithm, the ants construct solution (S) by using the following formula. Ant, located at object x1(i=1,….N), selects clusterj (j=1,….K) in probability Pij .

$$P_{ij} = \frac{\tau_{ij}[\eta_{ij}]^{\beta}[path_{if}]^{\alpha}}{\sum_{k=1}^{K} \tau_{ik}[\eta_{ik}]^{\beta}[path_{ik}]^{\alpha}}, j = 1,...., K \quad (8)$$

Where Pij is the probability distribution of object xi belonging to clusterj. $\eta_{ij} = 1/DWC(x_i, C_j)$ represents heuristic information value $DWC(x_i, C_j)$ is the DWC distance between object xi and the center of clusterj). β is the heuristic factor, indicating the relative importance of heuristic information. ij path denotes the number of excellent ants , which construct good solutions and group xi into the clusterj . If pathij is very large, we can speculate that building good solution must

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

group xi into the clusterj[13]. The Pij reflects that if pathij is very large, then xi is grouped into the clusterj with a high probability. By using this way a good solution has been developed rapidly.

4.2. Pheromone Update Rule

After each loop of the algorithm, i.e., when R(R≥5) ants have completed a solution, then the solution is sorted according to the clustering object function value in ascending order. Then, it get S sorted= $\{s_1^{'}, s_2^{'}, ..., s_R^{'}\}$, where $s_q^{'} = \{c_{q1}^{'}, c_{q2}^{'}, ...c_{qN}^{'}\}$, (q=1,….R)of which use the top 20% better solutions (S_best= $s_q^{''} \in S\_sorted, 1 \le q \le 20\%R \in Z\}$) to update the pheromone matrix. Pheromone updating formula is as follows:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \nabla\tau_{ij}(t)$$

$$\Delta\tau_{ij}(t) = \begin{cases} \sum_{s_q^{''} S\_best} \frac{Q}{DWC(x_i, Cs_q^{''} j)}, if c_q^{''} = cluster_j \\ 0, otherwise \end{cases} \quad (9)$$

$$\text{path}_{ij} = \begin{cases} path_{ij} + 1, is_q^{'} \in S\_best \wedge c_{qi}^{'} = cluster_j \\ path_{ij}, otherwise \end{cases}$$

Where ρ,0≤ρ≤1, is a user-defined parameter called evaporation coefficient, Q is a positive constant.

5. Fuzzy Enabled Clustering Scheme

5.1 Fuzzy Logic Concepts

Fuzzy Logic: It is based on “degrees of truth” rather than the usual “true or false” (1 or 0). It has been extended to handle the concept of partial truth. The truth values between “completely true” and “completely false”. Fuzzy logic uses the whole interval between 0 and 1 to describe human reasoning.

Fuzzy Set: An element either belongs or not belongs to the set.

The data-clustering scheme is designed by integrating ACO with DWC. The advantage of DWC is maintained the local consistency and global connectivity factors. The DWC and ACO enhanced using the fuzzy logic technique to improve the accuracy. The fuzzy enhancement is done in two stages, they are:

→ Distance estimation function is enhanced with fuzzy models to handle uneven data distributions.

→ The ant colony-clustering algorithm is enhanced with fuzzy relationship analysis model.

DOI: 10.7321/jscse.v2.n5.2

The novel method is designed with dynamic distance measure and enhanced with fuzzy enabled ant colony clustering models, which contains four stages. They are:

→ Fuzzification
→ Fuzzy based distance estimation
→ Clustering process
→ Cluster analysis phases.

The fuzzification phase is used to convert the attribute weight for each transaction into fuzzy sets. The distance and global connectivity are analyzed using the fuzzy weight values. The clustering process is done with the dynamic distance based ant colony clustering algorithm and fuzzy enabled ant colony clustering algorithm. Cluster accuracy is analyzed in the cluster analysis.

The attribute weight values for each transaction are passed into the fuzzification process. All the attribute weight values are converted into fuzzy weight values. The fuzzy weight conversion process updates the weight values with in a range of 0 to 1. The weight value distribution is not even in some data sets. The fuzzification process removes the overhead to calculate distance in uncertain data distributions and handles uneven transaction distribution. The distance measure estimates the distance with global connectivity factors. Fuzzy logic is applied to the distance estimation process. The dynamic distance based ant colony clustering algorithm is enhanced with fuzzy comparison for dynamic similarity analysis. All the transaction analysis is carried out with fuzzy enabled weight values. The cluster results are updated using the actual attribute weights. Fuzzy logic techniques are used to improve the distance estimation process. Global relationship is used in the system.

The comparison process is performed with fuzzy weights. The precision/recall and fitness measures are used in the cluster analysis process.

This progression has been classified into four processes.

→ Distance Analysis
→ Fuzzification Process
→ Ant Colony Clustering
→ Fuzzy Ant Colony Clustering

5.2 Distance Analysis

Distance analysis is performed to estimate transaction relevancy. Local and global distance estimation schemes are used in the system. Local distance is estimated with the current transaction information only. Global distance estimation uses the transaction details and support information

5.3. Fuzzification Process

Fuzzification comprises the process of transforming crisp values into grades of membership for linguistic terms of fuzzy sets. The membership function is used to associate a grade to each linguistic term [14].

DOI: 10.7321/jscse.v2.n5.2

Fuzzy weight is used for the distance estimation and also calculating Support value. Fuzzy model is used to assign weights in a range between 0 to 1. Transaction weights are converted into fuzzy based weights. First the data will be imported, then it will be cleaned, then the cleaned data will be converted into fuzzified values based on three ranges.

→ cz - Values lies between 0 to 4 (a=0, b=2, c=4) - if the data less than either 0 or 3 cz is set to zero – if the data lies between o to 2 cz will be updated using this formula ((data-a)/(b-a)) – if the data lies between 2 to 4 then the cz will be updated using this formula cz = ((c-data)/(c-b))

→ co - Values lies between 2 to 8 (a=2, b=5, c=8) - if the data less than either 4 or 8 co is set to zero - if the data lies between 2 to 5 co will be updated using this formula ((data-a)/(b-a)) - if the data lies between 5 to 8 then the co will be updated using this formula ((c-data)/(c-b)).

→ cb - Values lies between 6 to 10 (a=6, b=8, c=10) - if the data less than either 6 or 10 cb is set to zero - if the data lies between 6 to 8 cb will be updated using this formula ((data-a)/(b-a)) - if the data lies between 8 to 10 then the cb will be updated using this formula ((c-data)/(c-b)).

5.4. Ant Colony Clustering

The ant colony optimization algorithm (ACO) is a technique, can be applied to any optimization problems. It helps to find good and shortest path through pheromone trial updation. It is used to solve both static and dynamic optimization problem [10].

It uses its intelligent behavior to find the optimal path and contains metaheuristic optimization.

The ant colony optimization algorithm (ACO), is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. Initially proposed by Marco Dorigo in 1992. It was aiming to find optimal path between ant colony and food source through ant's path searching behavior [11].

Ant colony optimization technique is used for the clustering process. Here the transaction weights are used in the clustering process.

Algorithm:

1. Ant traverse around the colony to find the food source
2. After finding the food source it returns to nest.
3. While travelling it deposit some amount of pheromone.
4. The followers of the first ant follow the pheromones which left by the first ant.
5. This transaction will make strengthen the deposition of the pheromone.
6. This strengthens the route of the ant in mean time the amount of pheromone will evaporate in each traversal.
7. If there are two routes to reach the same food source the ant find the shortest route between food and nest with the help of pheromone updating.

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

5.5. Fuzzy Ant Colony Clustering

Clustering process is performed using fuzzy weights. Fuzzy weights based distance is used for the relevancy estimation. Global distance is used for the clustering process. Fuzzy relation is integrated with the ant colony clustering model is called FACO [12].

Algorithm:

1. Data set has to be imported.
2. Data cleaning must be done using precision and recall model.
3. The attribute support details will be calculated which consists of attribute value with corresponding count and support value.
4. The transaction support value will be calculated which consists of attribute name with corresponding attribute value and support value.
5. The support distance details for each data will be calculated.
6. The attribute value was converted into fuzzy values using the above mentioned formula in fuzzification process.
7. The original values are clustered with Ant colony clustering process.
8. In this process the cluster count was selected as per user needs then the details about the particular cluster were also seen.
9. Again the fuzzified values are clustered with Fuzzy Ant colony clustering process.
10. In this process the cluster count was selected as per user needs then the details about the
particular cluster were also seen.
11. In Fuzzy Ant colony clustering process the cluster accuracy has been improved and evaluated
with F-measure and entropy.

6. Experimental Analysis

The algorithms described in previous section are implemented using JAVA. For this excremental analysis Brest Cancer Data is used.

The cluster centroid optimization system is tested using the breast cancer diagnosis datasets. The dataset is taken from the UCI (University of California, Irwin) machine learning repository. Totally it contains 1000 data set and 11 attribute. It provides information about the breast cancer patient diagnosis information. The class information and associated symptom details are provided in the dataset. The dataset contains some noise records. Noise elimination process is performed on the data sets by using precision and recall methods. Clustering and rule mining operations can be tested using the dataset.

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

7. Evaluation Methods

The experimental evaluation is performed with the F-measure, entropy to evaluate the accuracy of the clustering algorithms. The F-measure measurement is used to represent the cluster accuracy information. Precision and recall values are used in the F-measure estimation process. Entropy is used to analyze the distance interval between clusters. The inter cluster intervals are analyzed using entropy measures.

The computational result for F-measure, Entropy is presented in Table 1. The performance analysis chart for F-measure is depicted in Fig 1, for entropy is depicted in Fig 2.

Table 1. shows the results of the F-measure analysis of clustering models.

Transaction	ACO		AFACO	
	F-Measure	Entropy	F-Measure	Entropy
200	0.718	0.875	0.927	0.942
400	0.739	0.886	0.941	0.956
600	0.758	0.898	0.956	0.963
800	0.779	0.908	0.972	0.984
1000	0.795	0.919	0.991	0.996

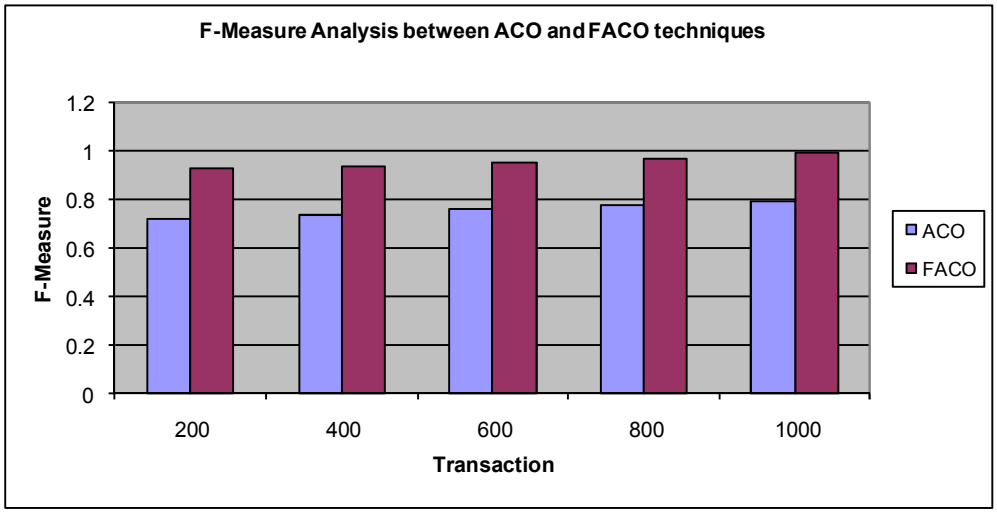

Figure 2 - F-measure Analysis of Clustering Models

International Journal of
Soft Computing And Software Engineering (JSCSE)
e-ISSN: 2251-7545
Vol.2,No.5, 2012
DOI: 10.7321/jscse.v2.n5.2

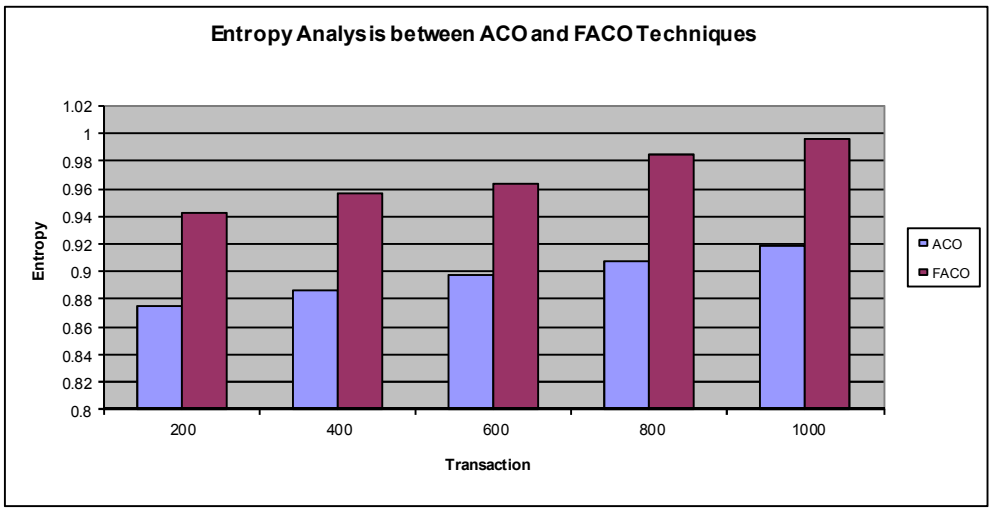

Figure 3 - Entropy analysis of Clustering Models

8. Conclusion

Based on the Euclidean distance between objects, the system uses data connectivity and an improved formula for calculating the distance named DWC. DWC reflects not only the local consistency but also the global connectivity between objects. It also overcomes the disadvantage of Euclidean distance in data clustering. Then, we improve the ant colony-clustering algorithm by using DWC and fuzzy logic concepts. Our experimental results on both synthetic and real world data sets show that the improved algorithm can discover clusters with arbitrary shape and is better than the clustering effect of earlier techniques. The limitation of this work is not reach to 100% accuracy. Further this work can be extended by incorporating rough set model to increase the accuracy clustering model for the breast cancer dataset.

References

[1] Jiawei Han and Micheline Kamber.Data Mining Concepts and Techniques, San Francisco: Morgan kaufmann, 2006, pp.383.
[2] Deneubourg JL, Goss S, et a1. “The dynamics of collective sorting: robot-like ant and ant-like robot,”. In: M eyer JA, Wilson SW ed. Proceedings first conference on simulation of adaptive behavior: from animals to animats. Cambridge, MA: MIT Press, 1991, pp.356–365.
[3] Shiyong Li, Baojiang Zhao. “Ant Colony Clustering Algorithm,” Measurement & Control, vol. 11, NO.15, 2007, pp.159–1592, 2007.15(11):1590–1592.
[4] Parag M. Kanade and Lawrence. Hall, “Fuzzy Ants and Clustering, IEEE Transactions on Systems,” man, and Cybernetics-part a: systems and humans, vol. 37, no. 5, September 2007, pp. 758–769.

DOI: 10.7321/jscse.v2.n5.2

[5] Xinbin Yang, Jinggao Sun and Dao Huang, "A New Clustering Method Based on Ant Colony Algorithm," Proceeding of the 4th World Congress on Intelligent Control and Automation June 10–14, 2002, pp.2222-2226.
[6] Jian Gao. "Cluster Analysis Based on Parallel Ant Colony Adaptive Algorithm," Computer Engineering and Application, vol. 25, 2003, pp.78–79, 2003.25:78–79.
[7] Yan YANG and Fan Jin, "Mohamed Kamel.Clustering Combination Based on Ant Colony Algorithm," Journal of the China Railway Socity, vol. 4, No. 26, 2004, pp.6–69.
[8] Maoguo Gong and Liefeng Bo. "Density-Sensitive Evolutionary Clustering," The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag Berlin Heidelberg ,2007, pp.507–514.
[9] Shanfei Li, Kewei Yang, Wei Huang, Yuejin Tan, "An Improved Ant-Colony Clustering Algorithm Based On The Innovational Distance Calculation Formula" Third International Conference on Knowledge Discovery and Data Mining, ISBN: 978-0-7695-3923-2, 2010
[10] Julia Handl and Joshua Knowles. "An Evolutionary Approach to Multiobjective Clustering," IEEE Transactions on Evolutionary Computation, vol. 11, no. 1, Feb.2007, pp.60.
[11] Miguel A.Sanz-Bobi and Mario Castro "IDSAI: A Distributed System for Intrusion Detection Based on Intelligent Agent" 5th International Conference on Internet Monitoring and Protection, IEEE, 2010.
[12] Amin Einipour "A Fuzzy-ACO Method for Detect Breast Cancer" Global Journal of Health Science October 2011.
[13] seyed saeed sadat noori, seyed ali sadat noori, seyed morteza lari baghal"Optimization of Routes in Mobile Ad hoc Networks using Artificial Neural Networks", International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 2, No. 4, pp. 36-50, 2012, Doi: 10.7321/jscse.v2.n4.4
[14] Mehdi Bahrami, Mohammad Bahrami, An overview to Software Architecture in Intrusion Detection System, International Journal of Soft Computing And Software Engineering (JSCSE), ISSN: 2251-7545, Vol.1,No.1, 2011.