Hybrid Ant-Fuzzy Approach for Data Clustering (AFC) in Distributed Environment

K. Sumangalairst [1] and Dr. S. Sathappan [2]

[1] Research Scholar, Kongunadu Arts & Science College, Coimbatore, Tamil Nadu, India
[2] Associate Professor, Erode Arts College, Erode, Tamil Nadu, India
[1]sumangala555@gmail.com
[2]devisathappan@yahoo.com

Abstract: Mining the relevant document from the distributed database is a Challenging task in the era of Big Data. This paper presents a hybrid approach of two different meta heuristics natural inspired algorithms namely Enhanced Ant Clustering Algorithm (EACA) and Fuzzy Clustering algorithm that deals with the uncertainty in data mining. The first one uses the Ant Colony metaphor that is one of the most recent nature - inspired meta - heuristics and the second one employs the fuzzy clustering approach. The proposed work aims to introduce the Fuzzy approach with Ant Based algorithm for both local and global that is inter clustering and intr-zone clustering in mining distributed databases and a new hybrid (Ant - Fuzzy) algorithm developed for data clustering under distributed environment. The proposed cluster based framework presents the important concepts of web mining and its various real time applications. The real time synthetic data and also the training data sets from the UCI repository were employed to evaluate the performance of the algorithms. The performance of the proposed work is compared with EACA, PACE and K-means algorithm in terms of accuracy with F-measure and Error rate with entropy measures.

Keywords: Ant Clustering, EACA, PACE, Fuzzy C_means, distributed clustering

1 Introduction

An era of big data Analytics, further databases appropriate on hand in the Internet and a growing number of online transactions. Data mining and distributed systems are leading the big data analytics..

Data mining is the computer-assisted technique that sequentially digs and analyzes the enormous sets of data to extract the information to meet certain requirements. In this web age numerous online sites are available basically they lie in two categories i.e., providing services to the people and doing smart business through that. The greater part of the community is used those sites based on their requirements. Some

sites display the queried results along with some other information. This kind of challenge was reduced using this proposed algorithm.

The Distributed database is a database in which storage devices are not close to a common processor. It may be stored in numerous storage places like secondary storage devices, other computers and cloud storages, etc., they are mostly dispersed over a network of organized computers and so on. This proposed approach has combined both mining and distributed environment with fuzzy based bio-inspired approach.

Ant-based clustering [5, 8] and categorization use two types of natural ant behavior i.e., while clustering, ants gather items to form heaps and while sorting, ants separate between different kinds of items and spatially arrange them according to their properties. In prior research work, EACA [1], k-means [3] and PACE [2]are used for Inter Zone (Global) Clustering. While using Ant based clustering always gives better results than others [12]. As per [11] in previous works were used various approaches for distributed clustering. In this work, Ant-Fuzzy Clustering Algorithm (AFC) is proposed with the use of Fuzzy Clustering for inter-zone clustering process. The Ant Colony Building[1] (ACB) and agglomeration[2] is modified by the Fuzzy Clustering algorithm in AFC, it regulates to trim down the few issues of the [1][2] and enrich performance of the process.

2 Related Work – Inter and Intra Zonal Clustering Algorithms

PACE: [5] Probabilistic Ant Based Clustering (PACE) algorithm [2] is based on the popular particle swarm algorithm of distributed data cluster [2,4,7]. Normally each and every family of ants are possessed an unique odor by them [6,7]. Using this odor only they might distinct from other family of ant. This behavior of ants is adopted in this algorithm to identify and form a group of ants carrying related data objects. In distributed database, the search keywords (data objects) are uniquely treated and identified from the databases of various data sites. The number of occurrences of keywords is computed using Hit-Ratio and the probability values based on Hit-Ratio are assigned to sites. The high order probability sites are considered and divided into larger zones. The ants are moved here and there without any restriction in their own area and they used to collect the various data objects (food) as they like. The data object to which they cluster around uniquely identifies the ant group and forms a group (Family). Then the each ant family begins to build their colony with collected data objects inside the zones based on the Ant Odor Identification Model [2]. The ants carry the data object to specific colony based on the picking and dropping probabilities (Ant Colony Building)[4] after forming the family *and* zones.. The colony is built similar to heap tree. Finally the heap trees formed by the ants are reordered or sorted to enable agglomeration. In PACE, the local & primary cluster was done by ACB and agglomeration was applied clustering of Inter-Zone.

EACA [2] Enhanced Ant Clustering Algorithm has included a special feature to cluster the distributed databases. This algorithm used the methodology of PACE with a modification, that is for applying ant clustering algorithm in intra and inter-zone clustering of data items. In EACA, Ant Colony Building [5,8] was used to cluster the

local data items within the zone and also outside the Zone for global clustering. An important features noticed in EACA , which gave better result than other algorithms in the form of accuracy and error rates.

3 Proposed Work

The proposed Ant-Fuzzy Clustering Algorithm combines the features of bio-inspired meta-heuristic algorithm and reality based fuzzy algorithm. Fuzzy is intimately connected with the concept of uncertainty. Here the most fundamental aspect of combining a Nature based algorithms with Fuzzy C-Means, especially ant clustering with Fuzzy c-Means for distributed databases are to find solution for the uncertainty involved among the cluster zones. Fuzzy c-Means allows one data may belong in two or more clusters based on minimization of the following objective function.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - C_j \right\|^2 1 \leq m < \infty \quad (1)$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional data, c_j is the cluster centre of d dimension data. Fuzzy partitioning is computed through an iterative optimization of the objective function with the membership updating u_{ij} and a cluster centers Cj using the formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\|x_i - c_k\|}{\|x_i - c_k\|} \right]^{\frac{2}{n-1}}} \quad (2)$$

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (3)$$

This iteration will stop when $(k+1) - u_{ij}(k) < \varepsilon$, where ε is a termination criterion between 0 and 1 and k is an iteration step.

The algorithm steps of proposed work for distributed database are as follows:

1, inquiring the databases

i. Initialize the counter value for Keywords

ii. Calculate the Thump-Ratio $Tr(r_m)$ of the keywords in various Database sites using the formula

$$Tr(r_m) = \sum_{d_m=1}^{d_m=m} \sum_{i=1}^{i=n} \frac{1}{k_i} \quad (4)$$

where k denotes keywords and n is number of keywords

2. Compute the Possibility of assorted database sites using the formula

$$P_r(d_m) = 1 - T_r(r_m) \quad (5)$$

Assign the ceiling of Zone formation for the database sites having higher probability.

3. Apply Ant Odor Identification Model to construct primary clusters and implement Fuzzy Ant Colony Building for intra zone clustering

c. Apply Fuzzy Clustering algorithm for inter zone clustering to group the clusters to single cluster that has high count of similar keywords

d. Clusters are validated using cluster validation measures.

After the completion of finite number of steps, the cluster results with the most relevant documents that retrieved from the distributed database.

4 Results and Discussion

The research work presents experimental results on synthetic and real-world data sets to investigate the properties of the proposed algorithm, and compare its effectiveness and scalability with related methods. The experimental setup is also carried out with the real time data as well as the benchmark datasets- “Iris” and “Wine” from UCI repository in order to assess the performance of the proposed algorithm. To evaluate the performance of algorithm, the datasets are permuted and randomly spread in the sites with a certain number of overlapping data.

The AFC has the enrichment of EACA [1] and [2] attention of ants is more just about highly apparent data object. Also sorting the heaps [1,2] of data are done to promulgate the grouping together of highly similar and most probable data. The evaluation methodology was inspired by [3,8].

The proposed algorithm compared with a popular k-means algorithm, PACE and EACA. Fig.1 describes the Ant based clustering of dataset and the different icons show the group of similar data objects.

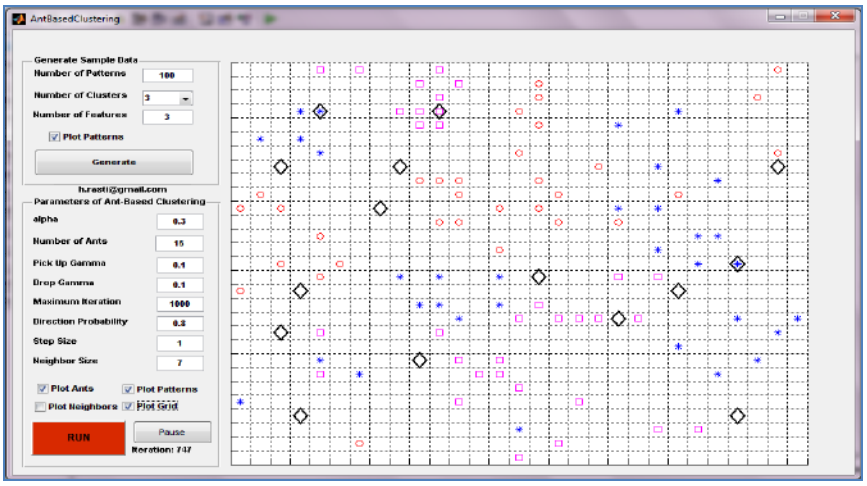

Fig. 1. Plotted Predicted data

Confusion Matrix is used to evaluate the performance of algorithm numerically. F-Measure computed from the confusion matrix which is adapted for comparing the clustering results and Table 1 show the result of F-Measure and Error rate of this work and others existing works.

Table 1: Comparison of F-measure **and** error rates for Iris and Wine datasets

Datasets	**Algorithms**	**F-Measure value**	**Mini-mum Errors**	**Maxi-mum Errors**	**Avg. Errors**
Iris	k-Means	0.8110	0.3	0.8	0.33
	PACE	0.8224	0.2	0.4	0.28
	EACA	0.8334	0	0.32	0.21
	Ant-Fuzzy	**0.9340**	0,12	0.31	**0.21**
Wine	k-Means	0.8217	0.55	0.83	0.57
	PACE	0.8771	0.3	0.45	0.31
	EACA	0.8995	0	0.36	0.29
	Ant-Fuzzy	**0.9015**	0.2	0.33	**0.24**

Table 1 depicts the outcome of the F-measure and error rate of other algorithms like K-Means, PACE, EACA and Ant-Fuzzy with Iris and Wine datasets. It is found that AFC performs better clustering than the existing K-Means and PACE whilst Ant-Fuzzy clustering results with 93% and 90% clustering accuracy for Iris and Wine datasets and proves its better performance than other methods.

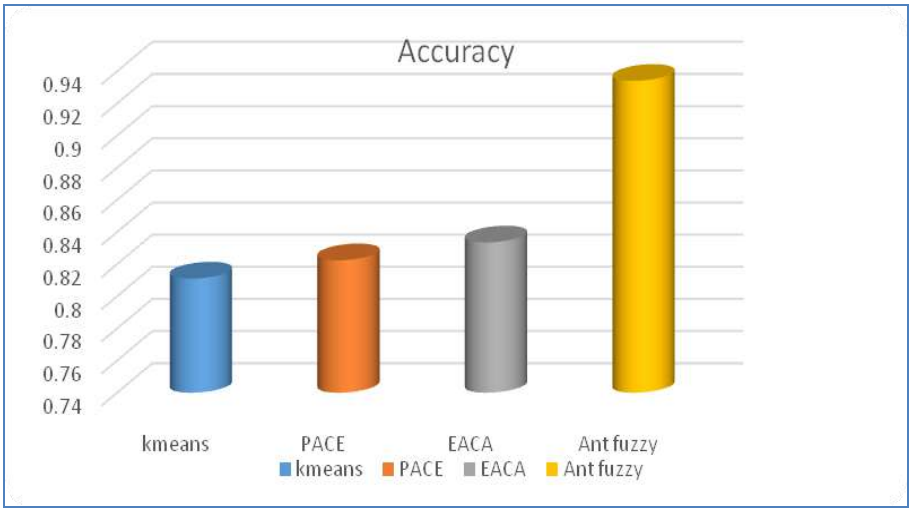

Fig. 2. Accuracy of clustering with "Iris" Dataset

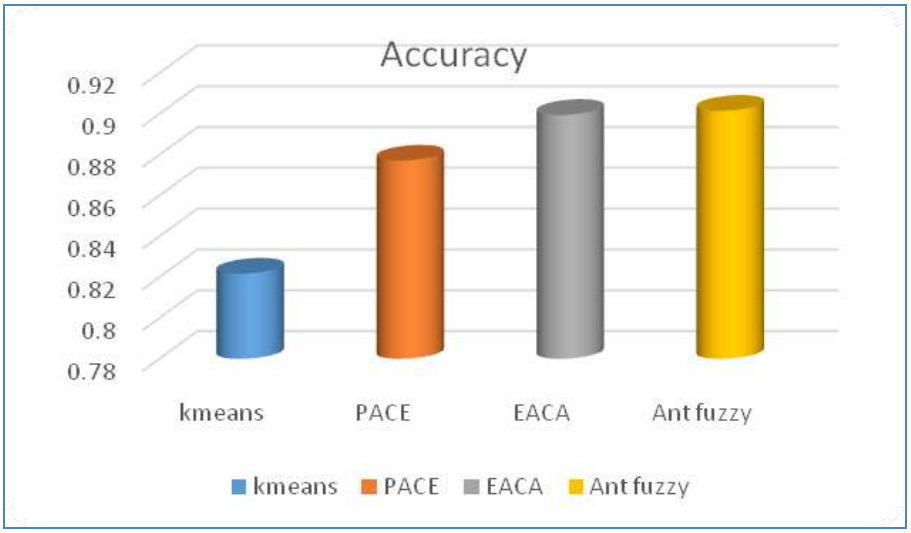

Fig. 3. Accuracy of clustering with "Wine" Dataset

Fig 2 and Fig 3 represent the clustering accuracy of proposed and existing algorithms. It is observed that the proposed Ant-Fuzzy based clustering algorithm cluster well for Iris dataset and performs better clustering than other algorithms for Wine Dataset. On averaging it is found that proposed clustering yields better result to retrieve relevant data from large distributed dataset.

The Table1 shows that the Clustering algorithm (Ant-Fuzzy) is quite successful in clustering the distributed databases with minimum error rate of 2% and the average error rate of proposed algorithm is 0.21 for Iris dataset is very less when compared to other existing algorithms namely k-Means, PACE and EACA. Similarly for wine data set also.

Fig.4 and Fig.5 shows the error rate of the proposed and existing algorithm for Iris and Wine datasets respectively. Which clearly demonstrate the proposed Ant-Fuzzy algorithm gives the less percentage of error than other existing algorithms.

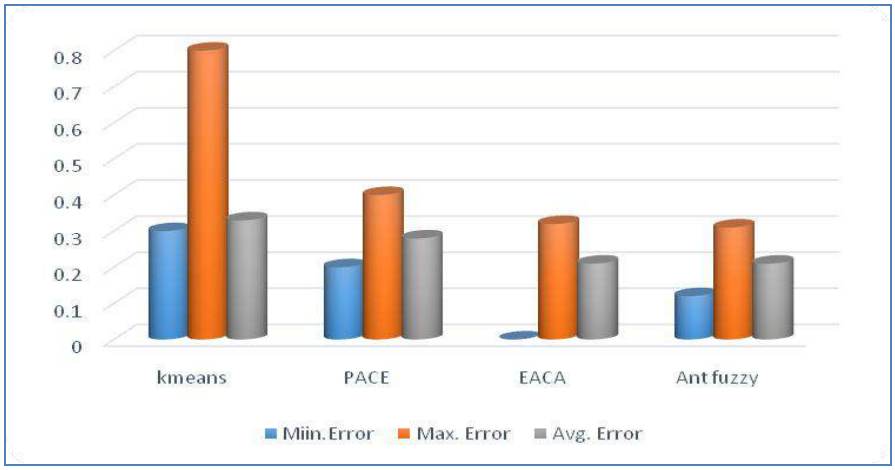

Fig. 4. Comparison of the Error rates with "Iris" Dataset

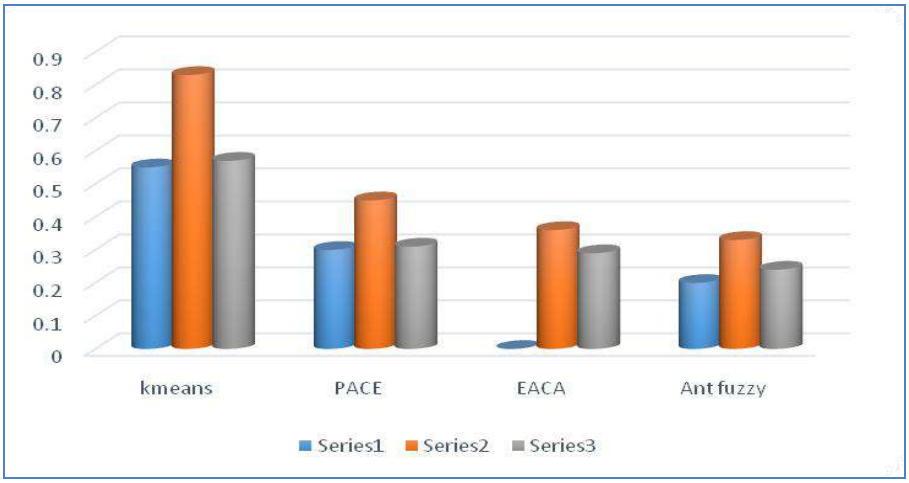

Fig. 5. Comparison of the Error rates on with "Wine" Dataset

Thus the above result confirms that the proposed algorithm produces better results than the existing algorithms.

5 Conclusion

The research work present an Ant based Fuzzy of clustering of the distributed databases. Here the Ant Clustering is combined with Fuzzy Approach the disseminated databases. The outcomes list out the AFC algorithm performs well. Error rare is reduced in each case of Ant-Fuzzy Clustering algorithm. The proposed algorithm possibly will assist the fresh and budding online sites to improve its performance while searching the data items from its distributed sites as well as local sites. The research, address a fuzzy based ant clustering algorithms and an overview of web usage mining applications attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web. This method can handle huge volume of heterogeneous data sets. The performances of this algorithm can also be tested for real commerce troubles. In future the Vector Quantization technique may be applied for Zone formation. In addition to AFC, further study can be combined and analyzed using Genetic Algorithm machine learning and AI techniques.

References

1. K. Sumangala, Enhanced Ant Clustering Algorithm, Proceedings of IEEE 4th International Conference on Computing, Communication & Network Technologies, India, July 2013.
2. R. Chandrasekar, Vivek Vijayakumar and T. Srinivasan. Probabilistic Ant based Clustering for Distributed Databases . In Proc. IEEE International Conference on Intelligent Systems 2006 , London, UK, September 2006.
3. E. Bonabeau, M. Dorigo, and G. Theraulaz, Swarm intelligence –from natural to artificial systems. Oxford University Press, New York, NY, 1999.
4. J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien, The dynamics of collective sorting: Robot-like ants and ant-like robots., In J.A. Meyer and S. W. Wilson, editors, Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats 1, pages 356–363, MIT Press, Cambridge, MA, 1991.
5. M. Dorigo, E. Bonabeau, and G. Theraulaz., Ant algorithms and stigmergy, Future Generation Computer Systems, 16(8):851–871, 2000.
6. M. Dorigo and G. Di Caro., Ant Colony Optimization: A new metaheuristic, In D. Corne, M. Dorigo, and F. Glover, editors, New Ideas in Optimization, pages 11–32.McGraw-Hill, London, UK, 1999.
7. J. Handl and B. Meyer., Improved ant-based clustering and sorting in a document retrieval interface, In Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature (PPSN VII), volume 2439 of LNCS, pages 913–923. Springer-Verlag, Berlin, Germany, 2002.

8. Handl, J., Knowles, J. and Dorigo, M .Ant-based clustering: a comparative study of Its relative performance with respect to k-means, average link and 1D-som. Technical Re Port TR/IRIDIA/2003-24. IRIDIA, Universite Libre de Bruxelles, Belgium, 2003.
9. Johnson E., Kargupta H ., Hierarchical clustering from distributed, heterogeneous data, In Zaki M. and Ho C., editors, Large-Scale Parallel KDD Systems, Lecture Notes in Computer Science, column 1759, pp. 221- 244. Springer-Verlag, 1999.
10. R. W. Klein, and R. C. Dubes, Experiments in projection and clustering by simulated annealing, Pattern Recogn. 22, 213-220, 1989. [16] C-Y Lee, and E. K. Anotonnson, Dynamic partitional using evolutionary strategies, In Proceedings of the Third Asia-Pacific conference on Simulated Evolution and Learning, Nagoya, Ja-pan, 2000.
11. Deepika Singh ; Anjana Gosain, A Comparative Analysis of Distributed Clustering Algorithms: A Survey, IEEE Digital Library 2013 .
12. Dhivya. N, Sumangala. K, A brief survey on Ant Based Clustering for Distributed Databases, International journal of Computer Sciences & Engineering, Volume-6, Issue-9, Page no. 540-544, Sep-2018.
13. James C. Bezdek, Robert Ehrlich, William Full. The Fuzzy c-Meams Clustering Algorithm, Computers & Geosciences Vol. 10, No.2-3,pp.191-203, 1984.