Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhanced Cat Clustering Approach

[1]K. Sumangala, [2]Dr. S. Sathappan
[1]Research Scholar, [2]Associate Professor
Research Department of Computer Science
Erode Arts College, Erode, India

Abstract— This paper proposes a novel approach for data mining named Enhanced Cat Clustering Algorithm (ECCA) and is one of the new swarm intelligence algorithms for finding the best global solution. This approach is inspired by the behaviour of cat or in other words it imitates the behaviour of cats. This work was tested with a training data set and the accuracy level was found to be improved when compared to other swarm intelligent Clustering Algorithms, especially ACO. Some times CSO clustering ACO clustering have appreciable error rate and the accuracy is less. For solving this problem and improving the accuracy and decreasing the error rate, a new improved approach for data mining is proposed in this work.

Keywords— clustering CSO, swarm intelligence, data mining

I. INTRODUCTION

CCA is a novel approach to cluster the data with the help of Swarm Intelligence[1]. Clustering is a useful technique for the discovery of data and patterns from the essential data. The goal of clustering is to discover the dense and sparse regions in a data set.

II. DATA MINING

Data mining (DM) is an interdisciplinary field and is used to extract knowledge from data sets (warehouse). DM is an. There are several data mining tasks, consisting of classification, clustering, association rule mining, regression etc. The initial pace in drawing a data mining algorithm is to describe which task the algorithm will tackle. In this paper Enhanced Cat Clustering Approach (ECCA) is for the clustering task of DM. So this article proposes a new approach (ECCA) of Cat Swarm Optimization (CSO) [3]. CSO is a contemporary stem of Swarm Intelligence.

III. SWARM INTELLIGENCE

Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial. The concept is employed in work on artificial intelligence. This was introduced by Gerardo Beni and Jing Wang in 1989, in the context of cellular robotic systems.[1]

SI systems consists natural population which is a simple means of interacting locally with one another and with their surroundings. This motivation arises from natural biological systems. Here the agents follow very simple rules and there is no centralized control structure because the interaction is dealt with local data. In global level there will be a intelligent behaviors to control the data. Examples in natural systems of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling etc.

IV. CAT SWARM OPTIMIZATION (CSO)

CSO is one of the innovative swarm intelligent algorithms for finding the finest ample resolution. Chu et al. [2] divided CSO algorithm into two sub- models based on two of the major behavioural traits of cats. These are termed as seeking mode and tracing mode. N number of cats is used for the clustering process. Every cat has its own position P, velocities for each dimension, and a fitness value. A flag is used to identify, whether the cat is in seeking mode or tracing mode.

4.1 Seeking Mode

This sub mode is used to model the cat during a period of resting at the same time being alert - looking around its environment for its next move [2] & [6]. Seeking mode has four essential factors, which are designated as follows: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC) and self position consideration (SPC). Seeking mode according to Chu et al. [3] is described below.

Step 1: Make j copies of the present position of cat k, where j = SMP. If the value of SPC is true, let j = (SMP - 1), then retain the present position as one of the candidates.

Step 2: For each copy, according to CDC, randomly plus or minus SRD percents the present values and replace the old ones.

Step 3: Calculate the fitness values (FS) of all candidate points.

Step 4: If all FS are not exactly equal, calculate the selecting probability of each candidate point by equation (1), otherwise set all the selecting probability of each candidate point be 1.

Pi = (FSi – FSb) / FSmax – Fsmin (1)

Step 5: Randomly pick the point to move to from the candidate points, and replace the position of cat k.

4.2 Tracing Mode

Tracing mode is the sub-model for modelling the case of the cat in tracing targets. Once a cat goes into tracing mode, it moves according to its' own velocities for each dimension [2] & [6].

The process of tracing mode can be described as follow:

Step1: Update the velocities for every dimension according to the following equation (2).

Vij = vij + r1 x c1 x (xbj – xij) (2)

Step2: Check if the velocities are in the range of maximum velocity. In case the new velocity is over range, it is set equal To the limit.

Step 3: Update the position of cat according to (3).

Xij = xij + vij (3)

Seeking and Tracing modes were applied as done by Budi Santosa [6] in his work.

V. ECCA

In this proposed approach two enhancements were made with the existing algorithm. The first enhancement was that the cat generally watches and gathers the required object and avoids unwanted object at any time. Second one is an initialization and updating the cluster which is automated with assist of enhanced k-means[6]. Enhanced k-means includes the random selection of clusters 'C' and calculate the distance between each data point with 'C'. The mean value of Ci is recalculated using the following formula to obtain the new cluster center.

$$CCi = (1/Ci) \sum Xi \quad (4)$$

This approach cluster the given data set at the moment, based on the characteristics of the cat this approach emphasizes the required group among the existing clusters.

Repeat the following steps 1-6 until stopping criteria is met

1. *Initialization of the clustering process (automated with enhanced k means)*
2. *Combine data based on the distance*
3. *Apply Seeking mode*
4. *Apply Tracing mode*
5. *Calculate SSE for Seeking and Tracing modes*
6. *Repeat*
   - *6.1 Calculate accuracy and error rate for each epoch*
   - *6.2 Compare the current accuracy and error rate with the previous one and get best cluster of each data*
7. *Choose the best cluster*

Algorithm (ECCA): Enhanced Cat Clustering Approach

VI. RESULT AND DISCUSSION

Table 1: Accuracy and Error rate

S.No	METRIC	EXISTING ALGORITHMS	ANT CLUSTERIG ALGORITHM	PROPOSED APPROACH
1.	ACCURACY	92.35	85.76	94.56
2.	ERROR RATE	0.2532	0.3792	0.2000

In this paper, simulation experiments are conducted in our aim is to reduce the error rate and increase the degree of accuracy. The following evaluation methods are used to measure the error rate and accuracy of this approach.

6.1 Evaluation Methods

6.1.1 F-Measure

F-measure combines the precision and recall concepts from information reclamation. We then calculate the recall and precision of that cluster for each class as

$$\text{Precision}(i, j) = (n_{ij} / n_j)\ i$$

$$\text{Recall}(i, j) = n_{ij} / n_i$$

where nij is the number of objects of class i that are in cluster j , nj is the number of objects in cluster j , and ni is the number of objects in class. The F - Measure of cluster given by the following equation

$$F(i,j) = (\text{Precision}(i, j) \cdot \text{Recall}(i, j)) / (\text{Precision}(i, j) + \text{Recall}(i, j))$$

The F-Measure values are within the interval [0, 1] and larger values indicate higher clustering quality.

6.1.2 Entropy

Entropy measures the purity of the clusters class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy increases.

To compute the entropy of a dataset, we need to calculate the class distribution of the objects in each cluster as follows

$$Ej = \sum Pij \quad \log (Pij)$$

Where the sum is taken over all classes the total entropy for a set of clusters is calculated as the weighted sum of entropies of all clusters, as shown in the next equation

$$E = \sum (nj / n)\ Ej$$

where nj is the size of cluster j , m is the number of clusters, and n is the total number of data points

In general the existing clustering and classification of algorithms [5,8] (SI based algorithms and others) provide the best outcome when compared [5,8] to other cluster and classification algorithm. But this proposed approach gives the clusters of given data as cited above and also highlights the user required data from the clustered data and is shown in figure5 and figure 7 clearly. It does not require the cluster initialization. Also the given data gets separated according to their classification automatically. So this approach gave best accuracy in all the classification and clustering of data. The following figures show the accuracy levels of data in each classification.

Table1 gives the accuracy and error rate of this proposed approach, which is also compared with the existing cso clustering and Ant clustering with constraints. This proposed approach gives best accuracy and decrease the error rate than the existing algorithms (cso clustering and Ant clustering with constraints). In the proposed approach the accuracy level and the error rate is calculated with the help the sum of bad clusters divided by the total number of data and also the accuracy was measured with the clustered data, outlier data and the total number of data items in the given data set. Some of the screen shots for the existing approach are in the following figures. Within the figures the figure 2 clearly shows the optimization level of the proposed work.

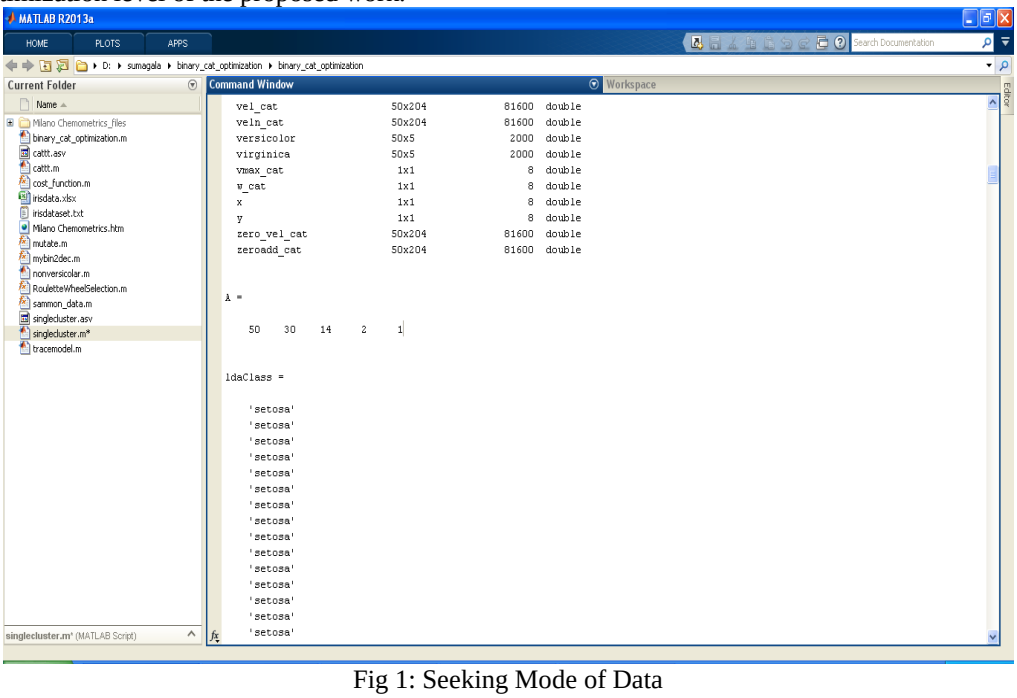

Fig 1: Seeking Mode of Data

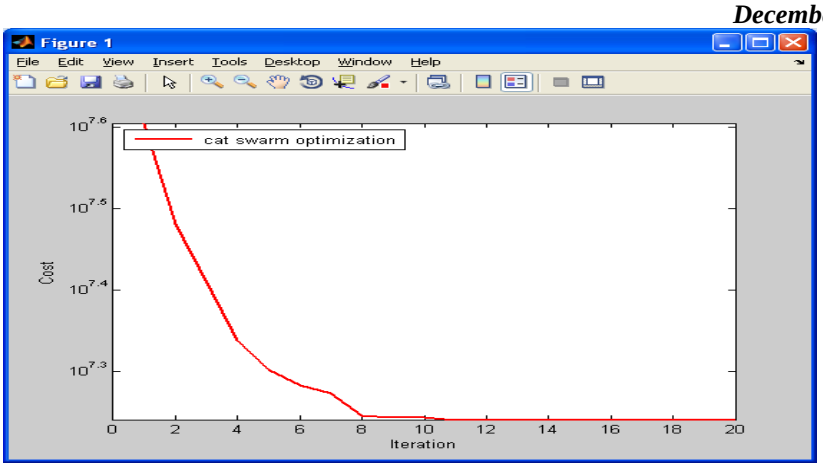

Fig.2 : Cost & Optimization level

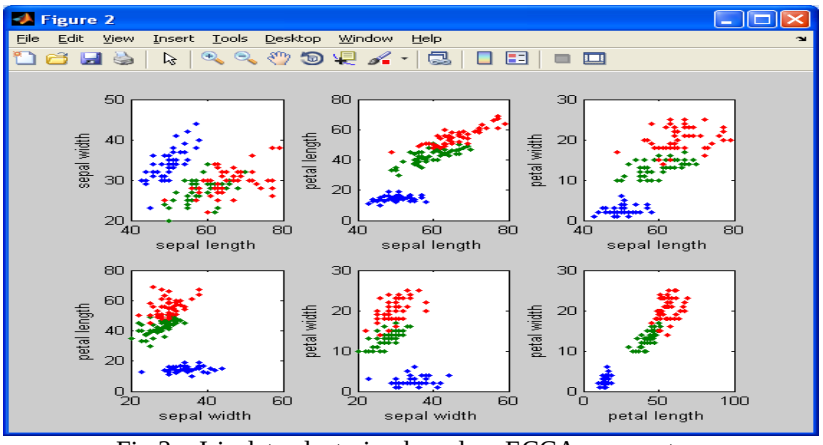

Fig.3 : Iris data clustering based on ECCA parameters

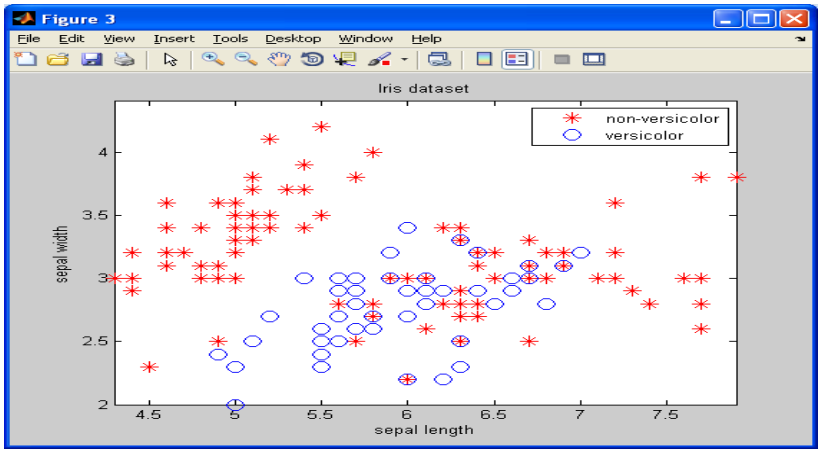

Fig. 4 : Total iris data set clustering all values based on sepal width and sepal length

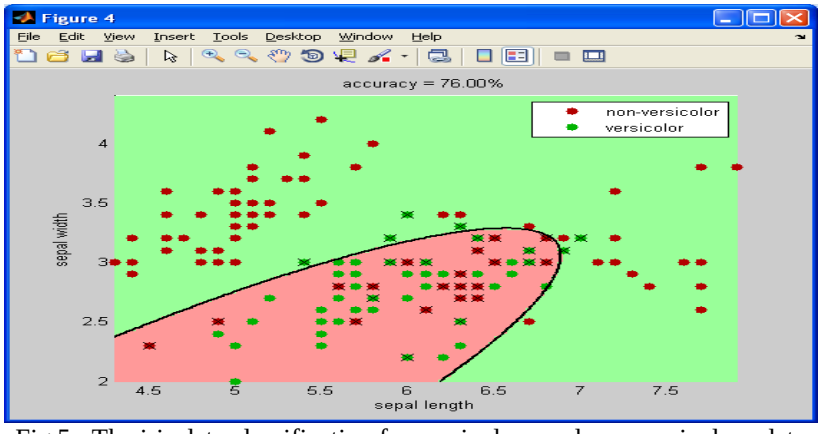

Fig.5 : The iris data classification for versicolour and non versicolour data

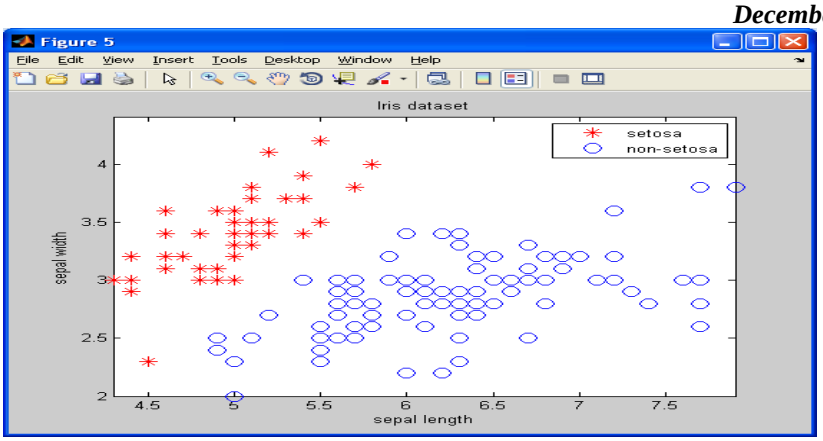

Fig. 6: Setosa and non setosa data separated and classified on iris data set

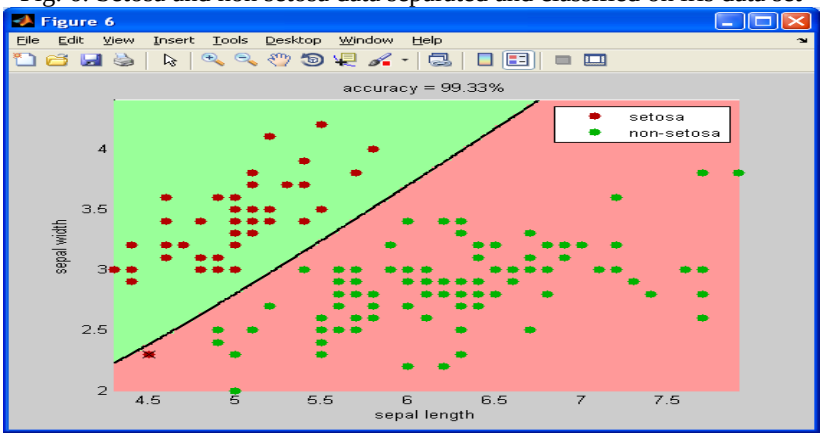

Fig .7: Setosa and non-setosa data classified 100 % accuracy based clustering

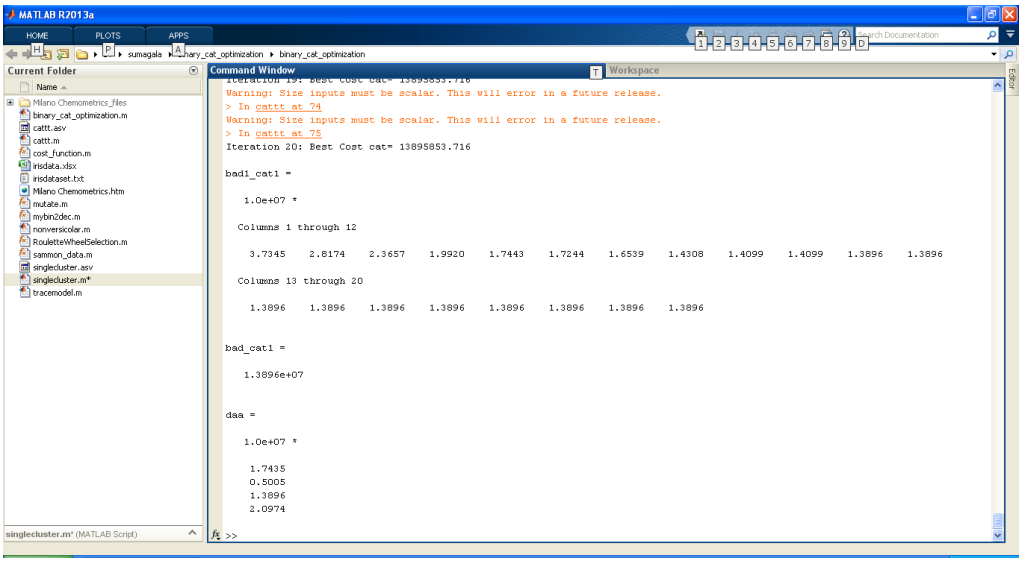

Fig 8: Cat optimization particles detection for iris data set.

VII. CONCLUSION

Clustering is the most researched topic of bio-inspired algorithms. Based on the literature review most of the clustering algorithms, the final clusters are based on the input value i.e., K value in basic K-means algorithms, number of ants in ant clustering algorithm. According to those values the centre may be varied. But in this approach does not need any initialization and centre calculations, these are automated. This paper has presented a focused assessment of accuracy and error rate for clustering. This comparison shows, CCA performs best clustering than the Ant family clustering algorithms based on the accuracy, error rate. This approach gives better outcome for local database clustering under the distributed data bases.

VII. FUTURE WORK

Normally a cat does not permit other anonymous cats within its region the same behaviour of cat will be introduced in this proposed approach for secured clustering.

REFERENCES

[1] Beni, G., Wang, J. Swarm Intelligence in Cellular Robotic Systems, Proceed. NATO Advanced Workshop on Robots and Biological Systems, Tuscany, Italy, June 26–30 (1989)

[2] Berkhin, P., 2002. "Survey of clustering data mining techniques." Technical Report, Accrue s/w Inc.

[3] Budi Santosa & Mirsa Kencana NingrumCat Swarm Optimization for Clustering, International Conference of Soft Computing and Pattern Recognition, 978-0-7695-3879-2/09, IEEE2009.

[4] S. C. Chu, P. W. Tsai, and J. S. Pan, "Cat Swarm Optimization," LNAI 4099, 3 (1), Berlin Heidelberg: Springer-Verlag, 2006, pp. 854 – 858

[5] I. Davidson, M. Ester and S.S. Ravi, "Clustering with constraints:

[6] Feasibility issues and the K-means algorithm", in proc. SIAM SDM 2005, Newport Beach, USA.

[7] Shi Yong; Zhang Ge; "Research on an improved algorithm for cluster analysis",International Conference on Consumer Electronics, Communications and Networks (CECNet), Pp. 598 – 601, 2011

[8] K. Sumangala & D. Vasanthi, An Enhanced Incremental Leader Ant Clustering with Constraints, Volume 42 - Number 20, IJCA 2012.

[9] K. Sumangala. Enhaced Ant Clustering Algorithm for Distributed Databases, Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference , ieee explorer.

[10] www.uci.repository

