A New Technique to Increase the Working Performance of the Ant Colony Optimization Algorithm

Reena Jindal, Samidha D.Sharma, Manoj Sharma

Abbstract- The DBSCALE [1] algorithm is a popular algorithm in Data Mining field as it has the ability to mine the noiseless arbitrary shape Clusters in an elegant way. Such meta-heuristic algorithms include Ant Colony Optimization Algorithms, Particle Swarm Optimizations and Genetic Algorithm has received increasing attention in recent years. Ant Colony Optimization (ACO) is a technique that was introduced in the early 1990's and it is inspired by the foraging behavior of ant colonies.

.This paper presents an application aiming to cluster a dataset with ACO-based optimization algorithm and to increase the working performance of colony optimization algorithm used for solving data-clustering problem, proposed two new techniques and shows the increase on the performance with the addition of these techniques [5]. We bring out a new clustering initialization algorithm which is scale-invariant to the scale factor. Instead of using the scale factor while the cluster initialization, in this research we determine the number and position of clusters according to the changes of cluster density with the division an agglomeration processes. Experimental results indicate that the proposed DBSCALE has a lower execution time cost than DBSCAN, and IDBSCAN clustering algorithms. IDBSCALE-ACO has a maximum deviation in clustering correctness rate of 95.0% and an error rate of deviation in noise data clustering of 2.62%.This algorithm is proposed to solve combinatorial optimization problem by using Ant Colony algorithm.

Keywords: DBSCALE, Ant Colony Optimization Algorithm, Clustering, Large Datasets.

I. INTRODUCTION

Data mining refers to the process of extracting or mining knowledge from large amounts of data. It involves the use of data analysis techniques to discover previously unknown, valid patterns and relationships in large data sets. Data mining tools perceive coming future trends and behaviors, allowing businesses to make positive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by showing tools typical of decision support systems. [1] Data mining tools can answer business questions that conventionally were too time consuming to resolve. They polish databases for hidden patterns, finding prophetic information that experts may miss because it lies outside their expectations. Data mining techniques are the outcome of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to find the way through their data in real time. [2] Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now satisfactorily mature[4] Data mining techniques can yield the benefits of computerization on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high piece parallel processing systems, they can examine massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze enormous quantities of data. Larger databases, in turn, yield improved predictions. [3] Perform partial analysis on local data at individual sites (nodes) and then send them to a central site (node) to generate global models by aggregating the local results. However, because the data is available in large-scale networks of autonomous data sources, a large number of nodes acting as information providers as well as information consumers into a dynamic information sharing system, the existing distributed clustering do not scale well. One of the biggest issues is to build good global models as local models do not contain enough information for the merging process. In this paper, we propose a new approach of distributed clustering. In our approach, basically we are trying to find the efficiency of clusters while reducing the noise as well as minimizing the outliers.

II. RELATED WORK

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. The data mining system may also integrate techniques from special data analysis, information retrieval, [5] pattern recognition, image analysis; signal processing, computer graphics, web technology, economics, business, bioinformatics, or psychology. Data mining system can be categorized according to various criteria,

i) Classification according to the kinds of databases mined.
ii) Classification according to the kinds of knowledge mined.
iii) Classification according to the kinds of techniques utilized.
iv) Classification according to the applications adapted.

Manuscript received on July 2013.

Reena Jindal, M.Tech Scholar, Dept. of Information Technology, NIIST, Bhopal, India

Dr.Samidha D.Sharma, HOD, Department of Information Technology NIIST, Bhopal, India

Prof.Manoj Sharma, Prof., Department of Computer Science and Engineering, NIIST, Bhopal, India

In our approach, we focus on the shape and the density of clusters created. The shape of a cluster is represented by its boundary points. Meanwhile its density can be represented by a mean density value or by a set of density values describing the density in various areas of the cluster. The algorithm to extract the boundaries from a cluster and the concept involved are The boundary of clusters as well as their density information will construct the reduction set φ that can be seen as the local model Mi at the site i in the system. [8] This local model is sent to the server in which global models will be built .The second step is for regenerating data points. In this step, each local dataset Di will be recovered by using its local model Mi. Obviously, the quality of recovering dataset depends on its local model. However, the regenerating methods have an important impact in the creation of a recovering dataset. We aim to build this new dataset so that the difference between it and the original local dataset Di is subject to minimization.
the data points.

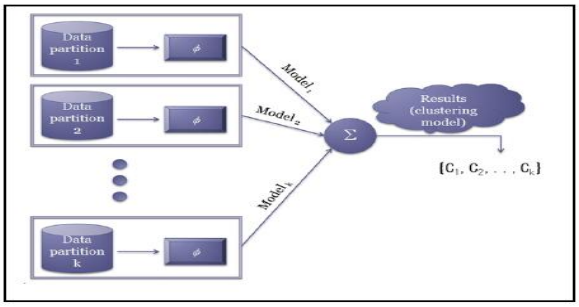


Figure 1. Distributed clustering model.

Clustering:

The process of grouping a set of physical or abstract object into class of similar object is called clustering. A cluster is a collection of data object that are similar to one another within a same cluster and different to objects in other clusters. A cluster of data objects can be treated collectively [9] as one group and so may be considered as a form of data compression.First partition the set of data into group based on data similarity and then assign labels to relatively small number of groups. Addition advantages of such a clustering based process are that it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization. The problem has been formulated in various ways in the machine learning, pattern recognition optimization and statistics literature. [10] The fundamental clustering problem is that of grouping together (clustering) data items that are similar to each other. The most general approach to clustering is to view it as a density based problem. Because of its wide application, several algorithms have been devised to solve the problem. Notable among these are the neural nets, DBSCAN and k-means. Clustering the data acts as a way to parameterize the data so that one does not have to deal with the entire data in later analysis, but only with these parameters [11] that describe the data. Sometimes clustering is also used to reduce the dimensionality of the data so as to make the analysis of the data simpler.The most widely used criterion for optimization is the distortion criterion. Each record is assigned to a single cluster and distortion is the average distance between a record and the corresponding cluster center. Thus this criterion minimizes the sum of the distances of each record from its corresponding center. K-means clustering is used to minimize the above-mentioned term by partitioning the data into k non-overlapping regions identified by their centers.

Density -based methods:-

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the “neighborhood “ exceeds some threshold: that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise and discover clusters of arbitrary shape. Density -based method s that grows clusters according to a density- based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions.

III. PROPOSED METHOD:ACO-DBSCALE

We proposed a new data point subset selection method for finding similarity matrix for clustering without alteration of density based clustering. The proposed data point subset selection method is based on ant colony optimization; ant colony optimization is very famous meta-heuristic function for searching/finding similarity of data. In this method, we have introduced continuity of ants for similar data points and dissimilar data points collect into next node. In this process, ACO finds optimal selection of data point subset. Suppose ants find data points of similarity in continuous root. Every ant of data points compares their property value according to initial data point set. When deciding data is noise and outlier, we should consider two factors: importance degree and easiness degree of noise and outliers. While walking ants deposit pheromone on the ground according to importance of the outlier and follow, in probability pheromone previously lay by other ants and the easiness degree of the noise.

Let D be the dataset and m be the number of ants, importance degree a_1, a_2,…., a_n is c_1, c_2 ,c_3 ……………..c_n, the appetency of solutions searched by two ants is defined as

$$App(i,\ j) = \frac{1}{ci - cj} \text{.................(1)}$$

where c_i and c_j is the importance of noise and outlier path. The concentration of the solution (1) is defined as

$$Con(i+\ j) = \frac{\delta i + \delta j}{m} \text{...........(2)}$$

where δ_i and δ_j is the number of ants whose appetency with other ants is bigger than α; α can be defined as m/10, then the incremented pheromone deposited by ants is

$$\Delta\tau_i(t) = Q.\beta_i / Con(i+j) \text{....................(3)}$$

where Q is constant.

Each level of pheromone modeled by means of a matrix τ where $\tau_{ij}(t)$ contains the level of pheromone deposited in the node i and j at time t, ant k in node i will select the next node j to visit with probability,

$$P_{ij}^{k}(t) = \{ \frac{[\tau ij(t)]^{\alpha}.[\eta ij]^{\beta}}{\sum_{l \in J_i^k} [\tau ij(t)]^{\alpha}.[\eta ij]^{\beta}} ifj \in J_i^k \text{ (1)...............(4)}$$

$0O$ therwise

where η_{ij} represents heuristic information about the problem which can be defined as the easiness of the path. The heuristic desirability of traversal and edge pheromone levels are combined to form the so-called probabilistic transition rule is given in equation (4), denoting the probability of an ant at data point i choosing to travel to data point j at time t.

Direct search in the best solution need global update rule applied as:

$$\tau(t+1) = (1-\rho).\tau_{ij}(t) + \rho.\Delta\tau_{ij} \text{....................(5)}$$

where $\rho(0 < \rho \leq 1)$ is a parameter that control the pheromone evaporation.

The steps of the proposed ACO based data preprocessing procedure for DBSCALE (ACO-DBSCALE) are as follows:

Step 1: Initialization of ants and degree of importance for the acceptance of data point selection: The appetency of solutions searched by two ants is defined as

$$App(i, j) = \frac{1}{ci-cj} \text{.................(1)}$$

where c_i and c_j is the importance of noise and outlier path.

Step 2: Find the acceptance solution on given parameter of degree of acceptance: The concentration of the solution (1) is defined as

$$Con(i+j) = \frac{\delta i + \delta j}{m} \text{...........(2)}$$

where δ_i and δ_j is the number of ants whose appetency with other ants is bigger than α; α can be defined as m/10, where m is the number of ants.

Step 3: Check the acceptancy of the data point and update the value of pheromone with amount of $\Delta\tau_i$: the incremented pheromone deposited by ants is

$$\Delta\tau_i(t) = Q.\beta_i / Con(i+j) \text{....................(3)}$$

where Q is constant.

Step 4: Generate the data point selection matrix after the increment of pheromone value and selected data points: Each level of pheromone modeled by means of a matrix τ where $\tau_{ij}(t)$ contains the level of pheromone deposited in the node i and j at time t, ant k in node i will select the next node j to visit with probability,

$$p_{ij}^{k}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum_{l \in J_i^k} [\tau_{il}(t)]^{\alpha} \cdot [\eta_{il}]^{\beta}} & if\ j \in J_i^k \quad (1) \\ 0 & otherwise \end{cases} \text{..........(4)}$$

where η_{ij} represents heuristic information about the problem which can be defined as the easiness of the path (η_{ij} is the heuristic desirability of choosing data point j when at data point i), J_i^k is the set of neighbor nodes of node i which have not yet been visited by the ant k. α > 0, β > 0 are two parameters that determine the relative importance of the pheromone value and heuristic information, and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i,j).

Step 5: Iterate and check data point matrix for processing of DBSCALE mapping: Direct search in the best solution need global update rule applied as:

$$\tau(t+1) = (1-\rho).\tau_{ij}(t) + \rho.\Delta\tau_{ij} \text{....................(5)}$$

where $\rho(0 < \rho \leq 1)$ is a parameter that control the pheromone evaporation.

Step 6: Finally, data point matrix is passed to DBSCALE algorithm for obtaining final clustering results.

IV. EXPERIMENTAL STUDIES

The clustering algorithm was implemented in the MATLAB Language in R2009a on a window XP operating System with a 2.0 GHz Intel Dual Core CPU, with 2G of RAM. The clustering correctness rate and noise filtering rate experiments were performed with eight different pattern datasets. With the aim of generating the optimal solutions of the presented ACO algorithm developed for solving data clustering problem and added two new techniques The experiments were performed on three algorithms in total. The absolute value (EPS) was fixed, then according to the pattern data density to adjust its Min Pts to conduct the experiment. The experiments measured (1) Execution time, (2) Clustering Correctness Rate and (3) Noise Filtering Rate or Error Rate, as Table I lists.

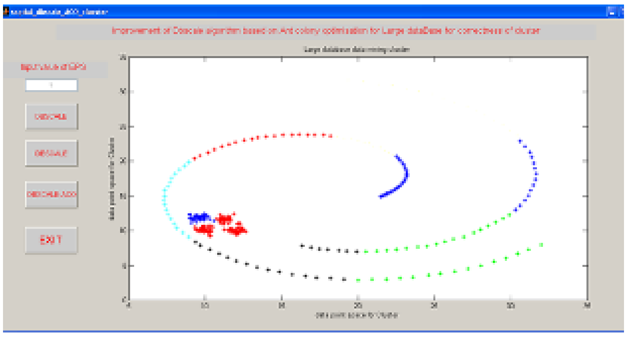

Improvement of Dbscale algorithm based on Ant colony optimisation for Large dataBase for correctness of cluster
Large database data mining cluster
Input value of EPS
1
DBSCALE
IDBSCALE
IDBSCALE-ACO
EXIT
data point space for Cluster

Figure 5.4.29 Shows that the DBSCALE of Sprial dataset for in input value 1

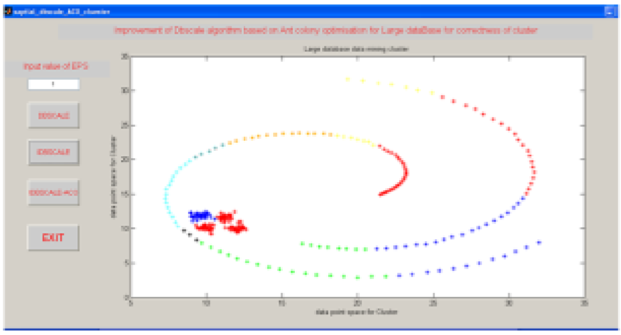

spatial_dbscale_ACO_cluster
Improvement of Dbscale algorithm based on Ant colony optimisation for Large dataBase for correctness of cluster
Large database data mining cluster
Input value of EPS
1
IDBSCALE
IDBSCALE
IDBSCALE-ACO
EXIT
5
10
15
20
25
30
35
data point space for Cluster

Figure 5.4.30 Shows that the IDBSCALE of Sprial dataset for in input value 1

Experimental results involving various pattern datasets comparing the existing IDBSCALE-ACO algorithms and those with neighborhood data search processing indicate a variation the clustering correctness rate of about 98.87% of Sprial Dataset, and the variation in the noise filtering rate never

exceeded 2.62%, revealing that the proposed algorithm does not affect the clustering quality and noise filtering capacity of the original algorithms, as Table I reveals. It is observed that the proposed DBSCALE-ACO algorithm is feasible for data clustering in data mining with large database.

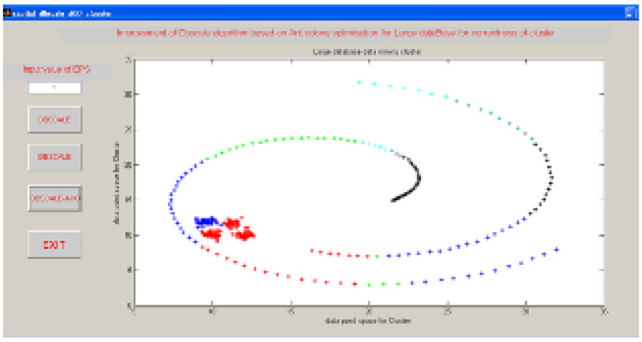
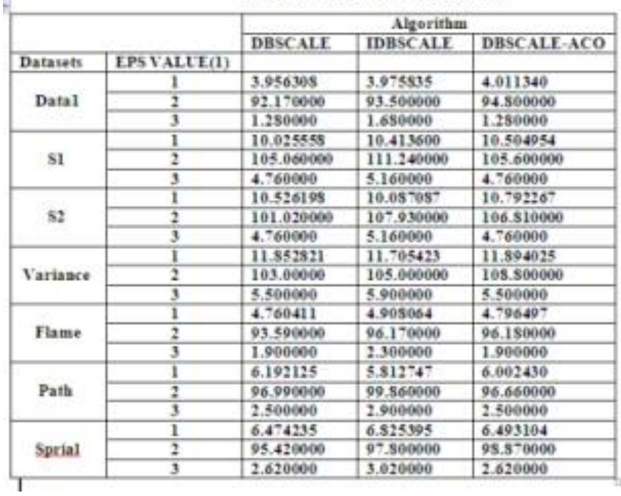


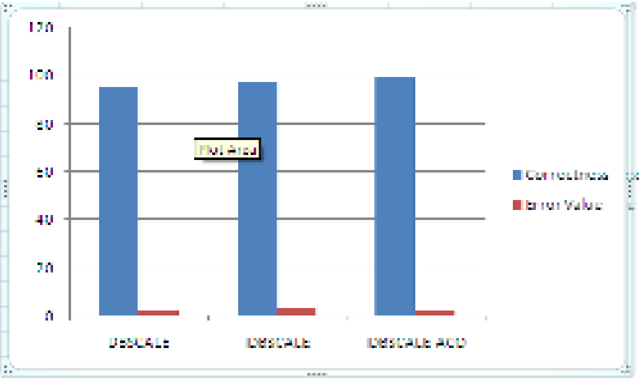
Figure 5.4.31 Shows that the IDBSCALE-ACO of Sprial dataset for in input value 1 of EPS the correctness

TABLE I

THE EXPERIMENTAL RESULTS



		Algorithm		
		DBSCALE	IDBSCALE	DBSCALE-ACO
Datasets	EPS VALUE(1)			
Data1	1	3.956308	3.975835	4.011340
	2	92.170000	93.500000	94.800000
	3	1.280000	1.680000	1.280000
S1	1	10.025558	10.413600	10.504954
	2	105.060000	111.240000	105.600000
	3	4.760000	5.160000	4.760000
S2	1	10.526198	10.087087	10.792267
	2	101.020000	107.930000	106.810000
	3	4.760000	5.160000	4.760000
Variance	1	11.852821	11.705423	11.894025
	2	103.00000	105.000000	108.800000
	3	5.500000	5.900000	5.500000
Flame	1	4.760411	4.908064	4.796497
	2	93.590000	96.170000	96.180000
	3	1.900000	2.300000	1.900000
Path	1	6.192125	5.812747	6.002430
	2	96.990000	99.560000	96.660000
	3	2.500000	2.900000	2.500000
Sprial	1	6.474235	6.825395	6.493104
	2	95.420000	97.800000	98.870000
	3	2.620000	3.020000	2.620000



V. CONCLUSION AND FUTURE SCOPE

In this Paper, We proposed new techniques to increase the working performance of the ant colony optimization algorithm the proposed techniques on an application program with the comparison of these three methods, it is shown that the proposed techniques increase the correctness of the reference IDBSCAN-ACO algorithm and the best results are derived from the third proposed technique. The ACO algorithm developed for solving the data clustering problem. New algorithm for IDBSCAN-ACO clustering is proposed which resourcefully overcome the major drawbacks viz. right number of cluster and initial seed (center point) problem. Proposed IDBSCAN-ACO clustering algorithm is based on two specific factors, threshold factor which initial decide the number of cluster and specific factor which merge the clusters according the similarity. The careful selection of threshold value and specific factor which control merging of clusters yields efficient algorithmic results.The area of future research includes that derive method for closing the appropriate threshold factor and this method should work for all dimensions i.e. various number of attribute it is also likely that the thresholds depend on the type of data. More research is needed for deriving exact threshold for this method.

REFERENCES

1. M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003
2. A. A. Freitas, S. H. Lavington, Mining very large databases with parallel processing. Dordrecht, The Netherlands, Kluwer Academic Publishers,1998
3. I. Foster, C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, Elsevier Press, 2004, pp. 593-620
4. T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting andrandomization. Machine Learning Vol.40, 2000, pp.139-158
5. Cheng-Fa Tsai, Chun-Yi Sung, "DBSCALE: An Efficient Density-Based Clustering Algorithm for Data Mining in Large Databases" (PACCS 2010) Second Pacific-Asia Conference on Circuits, Communications and System, 91201 Pingtung, Taiwan, October 2010.
6. R. Agrawal, J. C. Shafer, Parallel mining of association rules IEEE Transactions on Knowledge and Data Engineering, Vol 8., 1996, pp.962-969
7. E. Januzaj, H-P. Kriegel, M. Pfeifle, DBDC: Density-Based Distributed Clustering Proc. 9th Int. Conf. on Extending Database Technology(EDBT), Heraklion, Greece 2004, pp. 88-105
8. N-A. Le-Khac, L. Aouad, and M-T. Kechadi, A new approach for Distributed Density Based Clustering on Grid platform The 24th British National Conference on Databases (BNCOD'07), Springer LNCS 4587, July 3-5, 2007, Glasgow, UK. 2007
9. C. J. Merz, M. J. Pazzani. A principal components approach to combining regression estimates. Machine Learning Vol. 36, 1999, pp. 9-32
10. J. Kivinen, and H. Mannila, "The power of sampling in knowledgediscovery," Proceedings of the ACM SIGACT-SIGMOD-SIGART,Minneapolis, Minnesota, United States, May 24 - 27, 1994, pp.77-85
11. [11] K. Sayood, Introduction to Data Compression, 2nd Ed., MorganKaufmann, 2000