ISSN:0975-9646

Improved DBSCALE Algorithm by using Ant Colony Optimization

Reena Jindal[1,] Dr. Samidha D.Sharma[2,] Prof. Angad Singh[3]

[1] *M.Tech Scholar, Dept. of Information Technology, NIIST, Bhopal, India*
[2] *HOD, Department of Information Technology NIIST, Bhopal, India*
[3] *Prof., Department of Information Technology, NIIST, Bhopal, India*

Abstract- 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 than DBSCAN, and IDBSCAN clustering algorithms. IDBSCALE-ACO has a maximum deviation in clustering correctness rate. This algorithm is proposed to solve combinatorial optimization problem by using Ant Colony algorithm.

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

1 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.

2 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. Clustering is the process of grouping a set of data items (observations, feature vectors, or objects) into classes or clusters so that data items have high similarity when compared with one another within a cluster, but are very dissimilar to items in other clusters. Unlike in classification, the class label of each data item is previously unknown.

Clustering not only can act as a stand-alone tool, but also can serve as a pre-processing step for other algorithms which would then operate on the detected clusters. Those algorithms such as characterization and classification have wide applications in practice. All of them have attracted researchers from many disciplines, including statistics, machine learning, pattern recognition, databases, and data mining [12]. 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.

3 DBSCAN Clustering With Ant Colony Optimization

DBSCAN algorithm is a density-based clustering with high-quality, and applicable to geometry clustering of any shape and size, which can automatically determine the nuber of clusters, and separate clusters with environmental noise effectively. At first we need to introduce the related concepts of DBSCAN algorithm:

Definition 1: (Core Point) Point has at least MinPts objects in Eps neighborhood.

Definition 2: (Boundary Point) Point in the core object's Eps neighborhood, but not meets the requirements of core object.

Definition 3: (Directly Density-Reachable) A point p is directly density-reachable from a point q if

$$p \in NEps(q) \text{ and}$$
$$|NEps(q)| \geq MinPts$$

Definition 4: (Density-reachable) A point p is density-reachable from a point q if there is a

Chain of points $p_1,\ p_2,\ \ldots\ldots,\ p_n,\ \ p_1 = Q,\ p_n = p$ such that p_{i+1} is directly density-reachable from p_i.

DBSCAN algorithm efficiency depends on neighbor query and requires inspect each point in dataset by checking the neighborhood of each point to find cluster. If a point is a core point, then DBSCAN creates a cluster centered by this point and find directly density-reachable points. The algorithms time complexity is O(N log N), where n is the number of points in datasets.

ACO algorithm is inspired by ant's social behavior. Ants have no sight and are capable of finding the shortest route between a food source and their nest by chemical materials called pheromone that they leave when moving. ACO algorithm was firstly used for solving travelling salesman problem (TSP) and then has been successfully applied to a large number of difficult problems like the quadratic assignment problem (QAP), routing in telecommunication networks, graph coloring problems, scheduling, etc. This method is particularly attractive for feature selection as there seems to be no heuristic that can guide search to the optimal minimal subset every time on the other hand, if features are represented as a graph, ants will discover best feature combinations as they traverse the graph.

4 Proposed Solution

4/1 ALGORITHM

A new algorithm has been proposed in this paper to overcome the problem of the performance issue which exists in the density based clustering algorithms. In this algorithm first the data vectors are randomly scattered onto a two dimensional datasets. Ants (also called agents) are then randomly paced onto the two dimensional datasets. In each iteration step an ant searches its neighborhood and computes a probability of picking up a vector or of dropping down a vector. The ants moves randomly, search process can be speed up by traversing the data or guided by pheromone placed onto the datasets.

```
function [class,type]=dbscan(dataset,k,Eps)
 load location.mat
[m,n]=size(dataset);
if nargin<3 | isempty(Eps)
   [Eps]=epsilon(dataset,5);
end
dataset=[[1:m]' dataset];
[m,n]=size(dataset);
type=zeros(1,m);
no=1;
touched=zeros(m,1);
for i=1:m
    if touched(i)==0;
      ob=dataset(i,:);
      D=dist(ob(2:n),dataset(:,2:n));
      ind=find(D<=Eps);
      if length(ind)>1 & length(ind)<k+1
         type(i)=0;
         class(i)=0;
      end
      if length(ind)==1
         type(i)=-1;
         class(i)=-1;
         touched(i)=1;
      end
      if length(ind)>=k+1;
         type(i)=1;
        class(ind)=ones(length(ind),1)*madataset(no);

         while ~isempty(ind)
         ob=dataset(ind(1),:);
    touched(ind(1))=1;
     ind(1)=[];
     D=dist(ob(2:n),dataset(:,2:n));
          i1=find(D<=Eps);

              if length(i1)>1
                class(i1)=no;
                if length(i1)>=k+1;
              type(ob(1))=1;
            else
              type(ob(1))=0;
            end

            for i=1:length(i1)
              if touched(i1(i))==0
                touched(i1(i))=1;
                ind=[ind i1(i)];
                class(i1(i))=no;
              end
            end
          end
        end
        no=no+1;
      end
    end
end
```

```
i1=find(class==0);
class(i1)=-1;
type(i1)=-1;
function [Eps]=epsilon(dataset,k)
[m,n]=size(dataset);
Eps=((prod(dataset(dataset)min(dataset))*k*gamma(.5*n+1))/(m*
sqrt(pi.^n))).^(1/n);
function [D]=dist(i,dataset)

    function: [D]=dist(i,dataset)
[m,n]=size(dataset);
D=sqrt(sum((((ones(m,1)*i)-dataset).^2)'));

if n==1
  D=abs((ones(m,1)*i-dataset))';
end
```

4.2 FLOW GRAPH FOR PROPOSED WORK

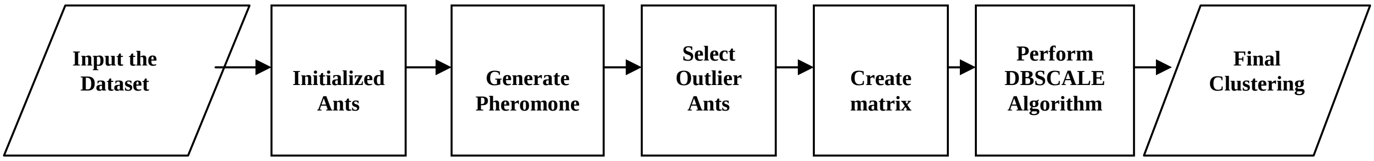

Figure I Flow Graph of Proposed Algorithm

5 EXPERIMENTAL EVALUATION

Experiments of the aim of generating the optimal solutions of the presented ACO algorithm developed for solving data clustering problem and added to new techniques. There is database consists of 12 data and it is stored in a text file. To the test of effectiveness of IDBSCALE-ACO algorithm we use four datasets. All datasets run with DBSCALE, IDBSCALE and IDBSCALE-ACO algorithm. The basic DBSCALE, IDBSCALE and proposed optimized DBSCAN algorithm are implemented in MATLAB 7.8.0 (R2009a) on Windows XP operating system and tested using two dimensional Datasets.. 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, as Table list I and II.

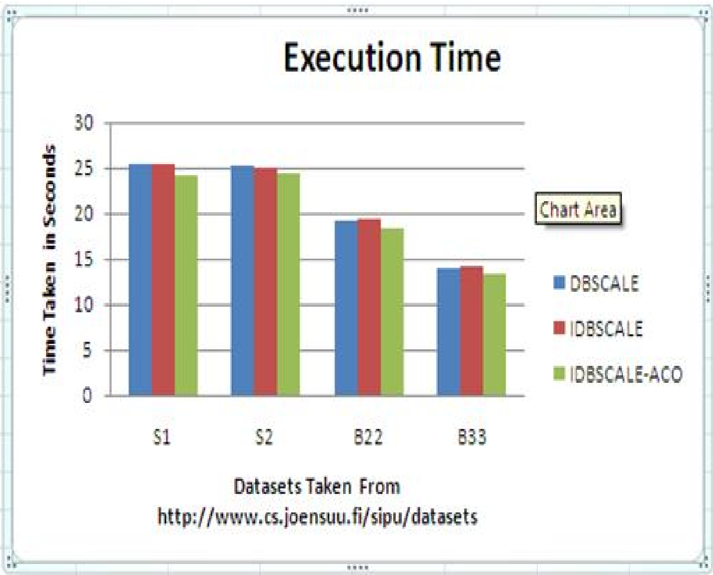

Figure I Time Elapsed of Proposed Algorithms in seconds

Dataset	DBSCALE	IDBSCALE	IDBSCALE-ACO
S1	25.5	25.5	24.2
S2	25.3	25.0	24.4
B22	19.1	19.3	18.4
B33	13.9	14.2	13.3

Table I Time Escaped of proposed algorithms in Seconds

Data set	DBSCALE	IDBSCALE	IDBSCALE-ACO
S1	72.14	74.63	78.41
S2	68.51	75.21	76.75
B22	87.85	88.90	89.47
B33	77.06	86.92	91.47

Table II Correctness of data in Proposed Algorithm

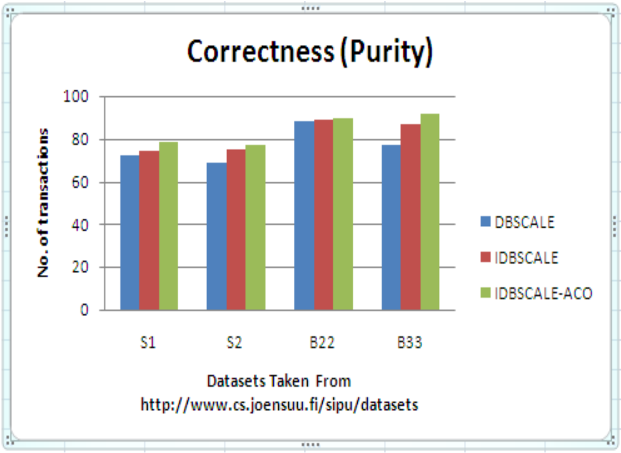

Figure II Correctness of data in Proposed Algorithm

6 CONCLUSION

In these 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 IDBSCALE-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 IDBSCALE-ACO clustering is proposed which resourcefully overcome the major drawbacks viz. Proposed IDBSCALE-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. Ant colony optimization has been and continuous to be a faithful designing effective combinatorial optimization solution algorithms. ACO one of the most successful in the meta heuristic area, there are a number of directions in which research on ant based clustering algorithms and clustering algorithm to real world applications. We also verified DBSCAN and proposed techniques on an application program with the comparison of these three methods, the proposed algorithm IDBSCALE-ACO is very well than DBSCAN and ACO clustering in terms of correctness measured by Rand index. One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data. There is lot of scope for the proposed Ant Colony Optimization Clustering algorithm in different application areas. Future work may address the issues involved in applying the algorithm in a particular application area.

REFERENCES

[1] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003.

[2] P-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006.

[3] Jean-Francois Laloux, Nhien-An Le-Khac, M-TaharKechadi, “Efficient Distributed Approach for density -Based Clustering”, IEEE 20thInternational Workshops on enabling Technologies, 2011.

[4] N-A. Le Khac, M. Whelan, and M-T. Kechadi, “Performance Evaluation of a Density-based clustering method for Reducing Very Large Spatio-temporal Datasets,” IEEE Sixth International Conference on Digital Information Management, (ICDIM’2011), Melbourne, Australia, September, 2011.

[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, Pingtung, Taiwan, October 2010.

[6] Michael Whelan1, Nhien-An Le-Khac2, M-Tahar Kechadi3, “Comparing Two Density-based clustering Methods for Reducing Very Large Spatio-temporal Dataset” IEEE International Conference on Belfield, Dublin 4, Ireland, November - 2011

[7] Nhien A Le Khac, Martin Bue, M-TaharKechadi, “Studying the Impact of Partition on Data reduction for Very Large Spatiotemporal Datasets” DBKDA 2011: The Third International Conference on Advances in Databases, Knowledge, and Data Applications. IARIA, 2011.

[8] Chunsheng Hua, Ryusuke Sagawa, Yasushi Yagi,” Scale-invariant density-based clustering initialization algorithm and its application” ISIR of Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, Japan.

[9] Lamia F. Ibrahim1,2 W. M. Minshawi2 Isra Y. Ekkab2 N. M. AL-Jurf2 A. S. Babrahim2 S. F. Al-halees2, “Enhancing the DBSCAN and Agglomerative Clustering Algorithms to Solve Network planning Problem” IEEE International Conference on Data Mining Workshops, Cairo University, Giza, EGYPT, November – 2009.

[10] Miroslav Bursa and LenkaLhotska, “Ant Colony Inspired Clustering in Biomedical Data processing” Czech Technical University in Prague, Technicka 2, Prague 6, Czech Republic.

[11] Mohd. Husain, Raj GaurangTiwarim, Anil Agrawal, Bineet Gupta “A New Ant Approach for Unraveling Data-Clustering and Data-Classification Setback”, International Journal of Computer Science and Application Issue 2010.

[12] Shu-Chuan Chu1, 3, John F. Roddick1, Che-Jen Su2, and Jeng-Shyang Pan2, 4, “Constrained Ant colony Optimization for Data Clustering”, C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004., Springer-Verlag Berlin Heidelberg 2004.