International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Improved the Correctness and Reduce the Error of DBSCAN using Ant Colony Optimization

Subodh Shrivastava[1], Brajesh Patel[2]

[1]M.E. Scholar, Department of CSE, SRIT, Jabalpur, India

[2]HOD, Department of CSE, SRIT, Jabalpur, India

Abstract: ***The processing and mining of spatial data is very challenging task in current research trend. The process of mining faced a lost due to diversity of data. The most part of spatial data contains a noise outlier, boundary point and core point. DBSCAN clustering faced a problem of noise of data and some boundary point of data. If the value of noise and boundary point reduces then we improved the correctness and performance of DBSACN algorithm. In this we improved the performance of DBSCAN algorithm using ANT colony optimization technique. Our experimental result shows that better performance instead of DBSCAN and IDBSCAN algorithm.***

Keywords: Clustering, DBSCAN, IDBSCAN, DENCLUE, OPTICS, CLIQUE.

1. Introduction

The process by which the knowledge is extracted from the huge amount of data is known as data mining. In this process the various data analysis techniques are used to discover the patterns which are valid and their relationships in large set of data [1]. The tools of data mining can distinguish the future trends and behaviors so that positive knowledge-driven decisions can be taken. Data mining provides the various automated decisions and analysis from the past events so that it will be useful for the future. These tools can provide the answers to the a variety of business questions which were time consuming [2].

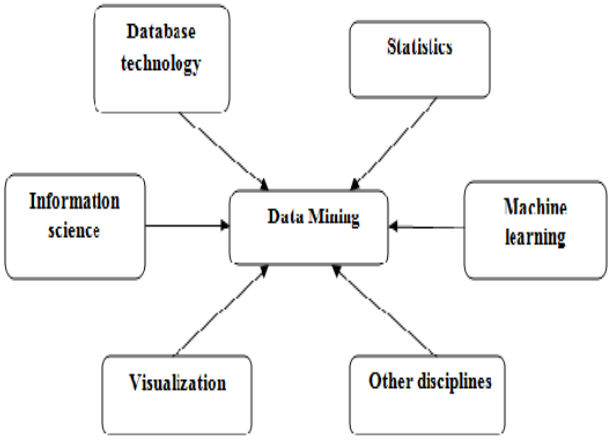

Figure 1. Data mining multiple disciplines.

2. Back Ground and Related Work

The process of organizing the similar objects in the same clusters and non similar objects in to different cluster is known as clustering. We can find Similarities between the objects by using characteristic value of object, and for finding the dissimilarity a distance metric is used. The K-means algorithm founds the clusters of convex shape but with DBSCAN algorithm we can find the clusters of arbitrary shape. DBSCAN algorithm is effective for analyzing complex spatial databases and large databases. DBSCAN need very large memory support and high dimensional data is not analyzed with it correctly but partitioning based DBSCAN can overcome these problems.

A. In DBSCAN algorithm each Cluster of radius (Eps) has to contain number of objects (MinPts). DBSCAN algorithm starts from an arbitrary point q, and retrieves density-reachable points from q . Eps and MinPts. If core point is q, then a cluster is formed and the point q and all points surrounded by q assigned to this new cluster. After this algorithm collects the points which are the in the distance Eps from the core points. This process continued till all number of the points processed. If any point is border point, and no points are reachable from that point then DBSCAN visits the subsequently point of the database [3]

B. MapReduce [4] is a parallel programming platform based on shared-nothing architectures. Ever since it was first introduced in 2003, MapReduce received great success due to its simplicity, scalability, and fault tolerance. Specifically, MapReduce provides users with readily usable programming interfaces while hiding the messy details for parallelism. Moreover, MapReduce divides a job into small tasks and materialize the intermediate results locally. As such, upon a node failure, only those tasks have to re-execute that are failed. Consequently, MapReduce algorithm may range up to thousands of commodity nodes where node failure is normal.

C. OPTICS [5] is a better form of DBSCAN algorithm which generates an order in which the objects needed to be processed. It then uses core-distance and reachability-distance in order to assign the each object a cluster membership. This order is generated through the reachability-distance and put in an ordered file. A cluster id is assigned to each object by using ordered file. Here Eps parameter plays an important role. By changing the Eps value, different structure of cluster is detected.

D. IDBSCAN [6] is an improvement of DBSCAN in terms of execution time. DBSCAN algorithm spends its major time in each object's region query. So, instead of expanding every object inside the region of core objects, IDBSCAN proposed

the expansion of those objects which are at the boundary region. The expansion of boundary objects would cover the objects which would are situated inside the region of core object, if had they been expanded. But it suffers from the limitations, similar to that of the DBSCAN.

E. LDBSCAN [7] is another algorithm which can detect different density based cluster. It uses the concept of LOF which represent the degree of outlierness and hence indicates whether the object is a core object or not. It then uses LRD [8] to assign an object in the direction of its matching cluster at some stage in the cluster expansion.

3. Design and Implementation

The process of mining of spatial data is very challenging task in the ground of data mining. The property of spatial data are very complex such as data contains maximum mixture contain such are called noise and distorted data in the process of gathering of spatial data. The part of noise and distorted data decrease the correctness of data during filtration process of data. The unfiltered and un-sampled data decrease the performance of spatial clustering technique such as DBSCAN and IDBSCAN. For the improvement of the correctness of data various authors used some optimization method for the lessening of noise and error of data. In this dissertation improved the correctness of DBSCAN algorithm via ANT colony optimization method. Ant colony optimization method processes the data in such a means it is artificial ant and pass through the technique of permanence. If the data point is dissimilar, the data predict as noise and reduces in part of data. The execution of process used as one process for the combination of DBSCAN and ant colony optimization method [9].

DBSCAN is an efficient process, but due to the noise proportion it is suffered by the data correctness problem. Its complexity increases with size of data set and then noise and outliers are too out of control. Ant Colony Optimization is very successful tool for finding quality data and that's the main reason to use it as a noise selection for DBSACN

We proposed a new noise selection cum reduction method for improvement of correctness of DBSCAN algorithm. Ant colony optimization is function for similar data searching/finding. In this method, we introduced ant's continuity for dissimilar noise and similar noise to collect into next node. ACO finds best possible selection of feature subset. If ants find similar noise in continuous root then every ant compares this initial feature set value. When data is outlier and noise, two factors are important easiness degree and degree of outliers and noise. While walking ants drop pheromone on the floor and at that time other ants also drop pheromone then according to importance of the outlier and trail, easiness degree of the noise [10].

If m is number of ants and D is the noise, importance degree will be a1, a2 ...an is c1, c2, c3 ….. cn, then the appetency solutions found by both ants is defined as

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

Here cj and ci and is the importance of outlier path and noise. The absorption of the solution (1) is define as

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

Here δi and δj are the quantity of ants whose appetency among other ants is more than α; here the α is defined as m/10, and the ants incremented pheromone deposited is :

$$\Delta\tau i = Q.\beta i / Con\ (i+j) \quad (3)$$

Here Q is constant.
every level of pheromone model by way of a matrix τ where τij(t) contain level of pheromone deposit in the i and j by time t, ant k can choose subsequently node j to trip with likelihood in the node i,

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

ηij is the heuristic information regarding the trouble. The heuristic popularity of the traversal and border pheromone levels is collected to form the hypothetical probabilistic evolution rule is given in the equation (4), denote the likelihood of ant at characteristic i choose to go to j at the time interval t.

Direct look for best solution need inclusive update rule applied as:

$$\tau\ (t+1) = (1-\rho).\ \tau ij\ (t) + \rho.\Delta\ \tau ij \quad (5)$$

here $\rho(0 < \rho \leq 1)$ is a factor that control the pheromone disappearance.

The steps of the planned ACO based information for DBSCAN follows:

Step 1: Ants Initialization and the grade of significance for the acceptance of characteristic selection: appetency of solution search by two ants is explained as

$$App\ (i, j) = \frac{1}{ci - cj} \quad (6)$$

here ci and cj is the significance of outlier path and noise.
Step 2: Find the approved solution on given constraint of degree of approval: The concentration of the answer explained as

$$Con\ (i+j) = \frac{\delta i + \delta j}{m} \quad (7)$$

here δi and δj is the quantity of ants whose appetency among other ants is higher than α; α can be defined as m/10, here m is the quantity of ants.

Step 3: Check the approval of the feature and revise the value of pheromone with quantity of Δτi: the incremented pheromone deposit by ants is

$$\Delta\tau i = Q.\beta i / Con(i+j) \quad (8)$$

here Q is the constant.

Step 4: Feature selection matrix is created subsequent to the increase of certain noise and pheromone value: level of pheromone noted by way of a matrix τ, τij (t) is the level of pheromone put down in the node i and the node j at time t

and ant k in the node i will prefer the next node j to trip with probability,

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\alpha}\cdot[\eta_{ij}]^{\beta}}{\sum_{i\in J_i^k}[\tau_{ii}(t)]^{\alpha}\cdot[\eta_{ii}]^{\beta}} & if\ j\in J_i^k\\ 0 & otherwise\end{cases} \quad (9)$$

here ηij represent heuristic information regarding the trouble which can be defined at the matching time as the easiness of path (ηij is the heuristic attractiveness of choosing feature j while at characteristic i), J_i^k is the adjacent nodes set of node i which are not previously visited by the ant k. The parameters $\alpha > 0$, $\beta > 0$ choose the comparative significance of the pheromone value and the heuristic information, and τij (t) is the quantity of implicit pheromone on border (i, j).

Step 5: Repeat and check the characteristic matrix for dispensation of SVM map: Direct look for the best answer need global update rule apply as:

$$\tau(t+1) = (1-\rho).\ \tau ij\ (t) + \rho.\Delta\ \tau ij \quad (10)$$

here $\rho(0 < \rho \leq 1)$ is a constraint that control the pheromone disappearance.

Step 6: At last, data matrix is passed to DBSCAN algorithm for obtaining cluster result.

4. Experimental Evaluation

To investigate the correctness of the proposed method for clustering dataset. We have performed some experimental task; all of these tasks perform in matlab 7.14.0 software and well famous UCI data set.

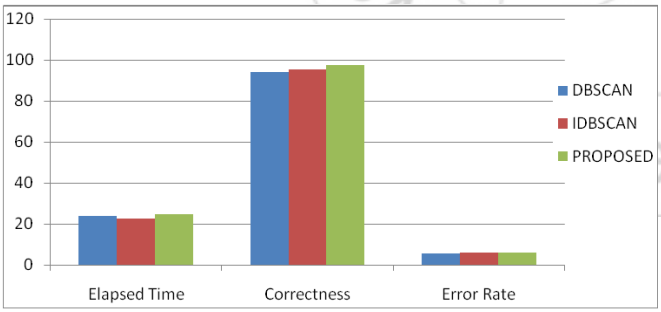

Figure 5.6.1: Comparative result

Above figure shows that the Comparative result graph for data set by using Proposed method, DBSCAN and IDBSCAN to find the value of Elapsed Time, Correctness and Error Rate. Finally we find that our proposed method gives better result and correctness than other method.

5. Conclusion

In this dissertation modified the DBSCAN algorithm for the improvement of correctness value of algorithm and also reduces the value of error for better generation and formation of cluster. The major problem related to DBSCAN algorithm is noise and outlier. The collected value of outlier and noise passes through the ANT colony optimization technique. The ANT colony optimization technique filters these values according to the value of continuous and discontinuous. The value of continuity and discontinuity shows that parameter selection and rejection according to the min point and radius. The value of radius increase the value of noise and boundary are decrease. 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.

References

[1] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491. B. Borah and D. K. Bhattacharyya, "An Improved Sampling-Based.
[2] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density based algorithm for discovering clusters in large spatial databases," in Knowledge Discovery and Data Mining, 1996.
[3] Huang Darong, Wang Peng, Grid-based DBSCAN Algorithm with Referential parameters
[4] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," in Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10.
[5] Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). Desity - based clustering over an evolving data stream with noise. In 2006 SIAM conference on data mining, Bethesda (pp. 328 – 339)
[6] G. Wei and H. Wu, "LD-BSCA:A Local Density Based Spatial Clustering Algoithm," in IEEE Symposium on Computational Intelligence and Data Mining. IEEE Computer Scoiety, 1999, pp. 291–298. ,
[7] Duan, Lian and Xu, Lida and Guo, Feng and Lee, Jun and Yan, Baopin "A local-density based spatial clustering algorithm with noise," Inf. Syst., vol. 32, pp. 978–986, November 2007.
[8] 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 14-16, 2011
[9] Dorigo M, Birattar i M, Stutzle T. Ant colony optimum: Artificial ants as a computational intelligence technique[J]. IEEE Computational Intelligence Magazine,2006, 1 (11) :28- 39.
[10] Junzhong Ji, Zhen Huang, Chunnian Liu, Qiguo Dai. An Ant Colony Algorithm Based on Multiple-Grain Representation for the Traveling Salesman Problems [J]. Computer Research and Development,2010.