Survey Paper On Different Techniques Of Measuring Efficiency Of Clustering

Reena Jindal[1,] Dr. Samidha D. Sharma[2,] Prof. Manoj Sharma[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 Computer Science and Engineering, NIIST, Bhopal, India

ABSTRACT

Data Mining is a process that uses technology to bridge the gap between data and logical decision making. The terminology itself provides a promising view of systematic data manipulation for extracting useful information and knowledge from high volume of data. Clustering is the process of organizing similar objects into the same clusters and dissimilar objects into different clusters. Similarities between objects are evaluated by using the attribute value of object, a distance metric is used for evaluating dissimilarities.

DBSCAN algorithm is attractive because it can find arbitrary shaped clusters with noisy outlier and require only two input parameters. DBSCAN algorithm is very effective for analyzing large and complex databases. DBSCAN need large volume of memory support and has difficulty with high dimensional data.

Ant Colony Optimization ACO) is a technique that is inspired by the foraging behavior of ant colonies. ACO algorithms have long been thought as generating high quality solutions for various problems. In this paper, we basically present the different techniques which can be used for measuring efficiency of clusters.

Keywords- Data Mining, Clustering, DBSCAN, DBSCALE, Ant Colony Clustering, Ant Colony Optimization.

INTRODUCTION

Data mining is a fast growing field in which clustering plays a very important role. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects [2]. Among the many algorithms proposed in the clustering field, DBSCAN is one of the most popular algorithms due to its high quality of noiseless output clusters. As the original DBSCAN algorithm Region Query function is very expensive factor in terms of time.

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e., it learns by observation rather than examples. There are no predefined class label exists for the date points.

Advantages

1. DBSCAN does not require you to know the number of clusters in the data priori, as opposed to k-means.
2. DBSCAN can find arbitrarily shaped clusters. It can even find clusters completely surrounded by a different cluster. Due to the mints parameter, the so –called single – link effect is reduced.
3. DBSCAN has a notion of noise.
4. DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database.

Disadvantages

1. DBSCAN can only result in a good clustering as good as its distance measure is in the function get neighbours. The most common distance metric used is the Euclidean distance measure. Especially for high – dimensional data, this distance metric can be rendered almost useless.

DBSCAN does not respond well to datasets with varying densities.

RELATED WORK

Classification scheme for clustering algorithms: - We classify well – known clustering algorithms according to different categorization schemes. Clustering algorithm can be classified along different, independent dimensions. One well – known dimension categorizes clustering methods according to the result they produce. Here, we can distinguish between hierarchical and partitioning clustering algorithms. Partitioning algorithms construct a flat partition of a database of an object into a set of clusters such that the objects in a cluster are more similar to each other than to object in different clusters.

We can classify clustering algorithms is from an algorithmic point of view. Here, we can distinguish between optimization based or distance based algorithms and density based algorithms. Distance based methods use the distances between the objects directly in order to optimize a global cluster criterion.

An overview of this classification scheme together with a number of important clustering algorithms is given in table:

In 2010, Cheng-Fa Tsai, Chun-Yi Sung developed a technique in which Clustered need not be input again when searching for neighbor-hood data points and the algorithm redefines eight Marked Boundary Objects to add expansion seeds according to far centrifugal force, which increases coverage. A novel clustering algorithm that incorporates neighbor searching and expansion seed selection into a density-based clustering algorithm.A new efficient DB SCALE clustering algorithm to solve the problem of data clustering for large databases. The presented novel clustering algorithm incorporates neighbor searching and expansion seed selection into a density-based clustering algorithm. Data Points that have been clustered need not be input again when searching for neighborhood data points and the algorithm redefines eight Marked Boundary Objects to add expansion seeds according to far centrifugal force, which increases coverage.

In 2011,Nhien-An Le-Khac, Michael Whelan, M-Tahar Kechadi, proposed a new scheme which deals (dranise) in the field of Data Mining work perform Distributed clustering technique as data-mining techniques to reduce very large spatio-temporal datasets into relevant subsets as well as to help visualization tools to effectively display the results. Cluster-based mining methods have proven to be successful at reducing the large size of raw data by retrieving its useful knowledge as representatives.A new approach for reducing large spatio-temporal datasets in the literature. This approach is based on the combination of density-based and graph - based clustering. In the future we intend to provide a more efficient algorithm to take into account the problem of large variation in density of datasets. Besides, we continue to carry out an extensive evaluation involving the analysis of more dimensions that have larger datasets than QCLOUD.

2.2 Chunsheng Hua, Ryusuke Sagawa, Yasushi Yagi,**" Scale-invariant density-based clustering initialization algorithm and its application**" deals (dranise) in the field of Data Mining work perform distributed clustering technique we determine the number and position of clusters according to the changes of cluster density with the division and agglomeration processes. During the division process, the initial cluster seeds are produced by a self-propagate method according to the density changes. The number of clusters is determined by agglomerating pair of RNN (reciprocal nearest neighbor) cluster seeds, when the density of newly merged cluster is increased. When no more cluster seeds can be merged any more, the remained number of cluster seeds is regarded as the real cluster number. We brought out a scale-invariant density-based clustering initialization algorithm. According to the density changes, a self propagate method is brought out in this algorithm to produce the cluster seeds at all the possible position. By defining the gap between clusters as the cluster boundary whose density is zero, the cluster detection can be achieved by checking the density changes before and after merging two RNN cluster seeds. Through experiments, this algorithm can be used as the initialization method for clustering, image segmentation, object tracking (like background subtraction), etc.

In 2010 Mohd. Husain, Raj Gaurang Tiwarim, Anil Agrawal, Bineet Gupta developed a new algorithm which deals (dranise) in the field of data mining work perform distributed clustering technique the data mining terminology, outlines the colony optimization algorithm which is used newly in data mining mostly aiming solve data-clustering and data-classification problems and developed from imitating the technique of real ants finding the short test way from their nests and the food source. an application aiming to cluster a data set with ant colony optimization algorithm and to increase the working performance of colony optimization algorithm used for solving data-clustering problem, proposes two new techniques and shows the increase on the performance with the addition of these suggested techniques**.** Two new techniques to increase the working performance of the ant colony optimization algorithm. We also verified ACO algorithm and proposed techniques on an application program With the comparison of these three methods, it is shown that the proposed techniques increase the performance of the reference ACO algorithm and the best results are derived from the second proposed technique. Consequently, our proposed two techniques markedly increased the success of the ACO algorithm developed for solving the data clustering problem.

In 1996 Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., in which deals (dranise) in the field of data mining work perform distributed clustering technique We present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shapes. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 benchmark. The results of our experiments demonstrate that (1) DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS, and that (2) DBSCAN outperforms CLARANS by a factor of more than 100 in terms of efficiency. Future research will have to consider the following issues: First, we have only considered point objects. Spatial databases, however, may also contain extended objects such as polygons. We have to develop a definition of the density in an Eps-neighborhood in polygon databases for generalizing DBSCAN. Second, applications of DBSCAN to high dimensional feature spaces should be investigated. In particular, the shape of the k-dist graph in such applications has to be explored.

COMPARISION OF DIFFERENT TECHNIQUES

1. DBSCALE: An Efficient Density-Based Clustering Algorithm for Data Mining in Large Databases**.**
Advantages:
1. A novel clustering algorithm that incorporates neighbor searching and expansion seed selection into a density-based clustering algorithm.
2. A new efficient DB SCALE clustering algorithm to solve the problem of data clustering for large databases.

2. Performance Evaluation of a Density-based Clustering Method for Reducing Very Large Spatiotemporal Dataset**.**
Advantages:
1. Cluster-based mining methods have proven to be successful at reducing the large size of raw data by retrieving its useful knowledge as representatives.
2. A new approach for reducing large spatio-temporal datasets in the literature.

3. Scale-invariant density-based clustering initialization algorithm and its application.
Advantages:
1. Clustering technique we determine the number and position of clusters according to the changes of cluster density with the division and agglomeration processes.
2. The initial cluster seeds are produced by a self-propagate method according to the density increase the working performance of the ant colony optimization algorithm.

4. A New Ant Approach for Unraveling Data-Clustering and Data-Classification Setback.\
Advantages:
1. Data mining mostly aiming solve data-clustering and data-classification problems and developed from imitating the technique of real ants finding the short test way from their nests and the food source.
2. an application aiming to cluster a data set with ant colony optimization algorithm and to increase the working performance of colony optimization algorithm used for solving data-clustering problem

5. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.
Advantages:
1. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it.
2. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 benchmark.

CONCLUSION

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.

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] 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.

[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 14-16, 2011.

[5]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, 567-0047, Japan.

[6] Mohd. Husain, Raj Gaurang Tiwarim, Anil Agrawal, Bineet Gupta **“**A New Ant Approach for Unr aveling Data-Clustering and Data-Classification Setback ”, International Journal of Computer Science and Application Issue 2010.

[7] M. Ester, H.-P. Kriegel, J. Xu, X. Sander, A Density-Based Algorithm for discovering clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, OR. AAAI Press (1996) 226-231.