Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Mining Actionable Clusters from High Dimensional Datasets

Preethi V,
PG Scholor,
Dept. of Information Technology,
SNS College of Engg.
Coimbatore, India

Suriya M,
Asst. Professor,
Dept. of Information Technology,
SNS College of Engg.
Coimbatore, India

Abstract—The datasets which are in the form of object-attribute-time format is referred to as three-dimensional (3D) data sets. Clustering these high dimensional (3D) data sets is a difficult task. So the subspace clustering method is applied to cluster the three-dimensional (3D) data sets. But finding the subspaces in the these three-dimensional (3D) dataset which is changing over time is really a difficult task. Sometimes this subspace clustering on three-dimensional (3D) data sets may produce the large number of arbitrary and spurious clusters. To cluster these three-dimensional many algorithms like MASC,TRICLUSTER,MIC,GS-Search and FCC is used now a days . But these algorithms allow the users to select the preferred objects as centroids. These algorithms are not the parallel one so they increases the time and space requirements which are needed to cluster the three-dimensional (3D) data sets. No optimal centroids have been chosen in the algorithms like MASC,TRICLUSTER,MIC,GS-Search and FCC to cluster the three-dimensional (3D) datasets. So for the first time in the proposed method the CPSO technique is introduced on the three-dimensional (3D) data sets to overcome all these drawbacks and CPSO clusters the three-dimensional (3D) datasets based on the optimal centroids and also it acts as the parallelization technique to tackle the space and time complexities.

Keywords—3D subspace clustering, singular value decomposition, numerical optimization technique, protein structural data analysis, financial and stock data analysis.

I. INTRODUCTION

Clustering is the task used to group the similar objects and because of its applications, clustering is popular with a large diversity of domains, such as geology, marketing, etc. Across the years, the tremendous amount of growth in the data has created a lot of high-dimensional data sets in these domains. As a consequence, the deviation between any of the two objects becomes same or similar in the high dimensional data, which reduces the meaning of the cluster. [2]

A technique introduced to handle this high dimensional dataset[3] is by clustering the dataset in its subspaces, and so the objects lie in a group is enough to be similar to a subset of attributes called subspace, rather than of living similar over the entire set of attributes called full space. Through SVD technique the subspace clustering will be made. But most the high-dimensional datasets in the domain like stock market can possibly change over time. So the handling of the dataset which is changing over time is really a difficult task. The data which is changing over time is referred to as three-dimensional (3D) dataset. These three-dimensional (3D) data sets[14] can be generally stated in the form of object-attribute-time, for instance the stock-ratio-year data in the finance area, and the residues-position-time protein structural data in the biological science, among others .In such data sets, discovering subspace clusters per timestamp may produce many spurious and arbitrary clusters, thus it is worthy to detect clusters that remain same in the database across the specified amount of time period. These three-dimensional (3D) dataset clustering is a difficult task. Even though many algorithms handle this issue, most of them are inadequate and not efficient to cluster three dimensional datasets.[18]

Space and time complexity problem also rises while clustering very large amount of data sets that contain large numbers of records with high dimensions is considered. This is a very important issue now a days. Examples are the clustering of profile pages in social networks, Bioinformatics applications, and article clustering of big libraries. Most sequential clustering algorithms suffers from the problem that they do not scale with larger sizes of data sets, and most of them are computationally expensive in memory space and time complexities. For these reasons, the parallelization of the data clustering algorithms is paramount in order to deal with large scale data. To develop a good parallel clustering algorithm that takes big data into consideration, the algorithm should be efficient, scalable and obtain high quality clusters.[11]

So definitely an algorithm needed to handle the dataset which is changing over time (3D dataset) and to reduce the time and space complexity while clustering the data have more timestamp and high dimensions. So the MapReduce method is introduced with CPSO algorithm which can handle the data that is changing over time based on the optimal centroid value and the parallelization will be employed by this MapReduce CPSO to reduce the time and space complexity.

II. LITERATURE SURVEY

A. MASC

MASC (Mining Actionable Subspace Clusters) is a method which is proposed to mine actionable subspace clusters from sequential data. MASC uses the subspace concept with high and correlated utilities. To efficiently mine 3D subspace clusters, MASC (Mining Actionable Subspace Clusters) is used, which is a hybrid of numerical optimization, principal component analysis and frequent itemset mining. The MASC algorithm flatten the continuous valued dataset into a dataset with a single timestamp. A wide range of experiments is conducted to demonstrate the actionability of the clusters and the robustness of the framework MASC. But MASC cannot generate true 3D subspace clusters. MASC can work well only on 2D datasets rather than 3D datasets. In MAC it is hard to find clusters in the dataset which has more number of timestamps.[11]

B. TRICLUSTER

TRICLUSTER [9] is the algorithm which is used to mine the 3D subspace clusters with the concept of subspace in all three dimensions. Its clusters are highly flexible as users can use different homogeneity functions such as distance, shifting, and scaling functions. Users are required to set thresholds on the parameters of these homogeneity functions and clusters that satisfy these thresholds are mined.

TRICLUSTER, along with most of the subspace clustering algorithms, are parameter based (clusters that satisfied the parameters are mined), and their results are sensitive to the parameters. In general, it is difficult to set the correct parameters, as they are not semantically meaningful to users. For example, the distance threshold [1], [4] is a parameter that is difficult to set , at any distance threshold setting, different users can perceive its degree of homogeneity differently. Moreover, at certain settings, it is possible that a large number of clusters will be mined.

C. MIC

Algorithm MIC [16] proposed mining significant 3D subspace clusters in a parameter insensitive way. MIC(Mining Interesting subspace Clusters) uses the concept of correlated 3D subspace clusters(CSC). Significant clusters are intrinsically prominent in the data, and they are usually small in numbers. The algorithm MIC uses the concept of significance, but they focus on mining interesting subspaces [5] or significant subspaces [10], and not on the mining of subspace clusters. Both TRICLUSTER and MIC do not allow incorporation of domain knowledge into their clusters, and their clusters are not actionable.MIC can only mine 3D subspace clusters which are the subsets of attributes and subsets of timestamps.[12]

D. GS-Search

GS-Search algorithm deals with the concept of n-ary relations which receives attention in many different fields, for instance biology, web mining, and social studies. In the basic setting, there are n sets of instances, and each observation associates n instances, one from each set. A common approach to explore these n-way data is the search for n-set patterns and the n-way equivalent of item sets. More precisely, an n-set pattern consists of specific subsets of the n instance sets such that all possible associations between the corresponding instances are observed in the data. In contrast, traditional item set mining approaches consider only two-way data, namely items versus transactions. The n-set patterns provide a higher-level view of the data, revealing associative relationships between groups of instances. Thus GS-Search converts 3D dataset[6] into single timestamp datasets. But GS-Search algorithms cannot handle the datasets that have larger timestamps.

E. FCC

Frequent Closed Cube (FCC) is a method which generalizes the notion of 2D frequent closed pattern to 3D context. Two novel algorithms to mine FCCs from 3D datasets are introduced. The first scheme is a Representative Slice Mining (RSM) framework that can be used to extend existing 2D FCP mining algorithms for FCC mining. The second technique, called CubeMiner, is a novel algorithm that operates on the 3D space directly. In 3D context the frequent closed pattern is referred as frequent closed cube (FCC). Even in the traditional 'market-basket' analysis, it is not uncommon to have consumer information on a number of dimensions.[8]

The problem of mining FCC from 3D datasets is solved by RSM. First, the notion of FCC is introduced and formally it is defined. Second, two approaches to mine FCCs is proposed. The first approach is a three-phase framework, called Representative Slice Mining algorithm (RSM) that exploits 2D FCP mining algorithms to mine FCCs. Even though it handles both 2D and 3D datasets it is not efficient for 3D datasets.RSM[2] performs best when one of the dimensions is small. FCC does not allow the incorporation of domain knowledge which is the main drawback of Frequent Closed Cube (FCC) algorithm.

TABLE I: SUMMARIZED RESULTS OF ALGORITHMS

	MASC [11]	**TRICLUSTER [9]**	**MIC [12]**	**GS-Search [6]**	**FCC [8]**
Domain Knowledge Incorporation	YES				YES
Parameter Insensitive		YES	YES		
Actionable	YES			YES	
Generates 3D Subspaces	YES				

III. PROPOSED METHOD

In proposed method the Mapreduce particle swarm optimization is used which uses the concept of centroids, to cluster the 3D datasets. In the MapReduce-CPSO algorithm, the clustering task is considered as an optimization technique to obtain the best and optimal clustering result based on the optimal centroid value. The optimal solution is obtained by calculating the distance between the data points and the centroid. The MapReduce-CPSO is similar to the K-means clustering algorithm. In CPSO the fitness value is calculated then based on fitness value the centroid value is updated. But in MapReduce-CPSO particle's velocity used to update the centroid value. In MR-CPSO the particles contain the information which is used to accelerate the clustering task. This MapReduce CPSO can work well with the increasing data sizes which is used to increase the cluster quality with minimal time and space requirement.

IV. CONCLUSION

This paper surveys the techniques of 3D subspace clustering. In earlier days many algorithms was used to cluster the three dimensional datasets. But most of them was inadequate to cluster three dimensional datasets. In this survey paper, many algorithms for clustering 3D datasets are reviewed. The algorithm called MASC) do not generate true three dimensional subspace clusters but MASC[11] can cluster 2D dataset very efficiently. The algorithms like Tricluster[9] and GS-search[6] requires users to set parameters in which results mainly depends on the tuned parameters. The algorithms like MIC[12] and FCC[8] do not allow the incorporation of domain knowledge into clusters and both of them are not actionable.

Our work is to propose new algorithm to tackle all the difficulties of above mentioned algorithms. In the proposed method the algorithm called MapReduce CPSO is applied to the large amount datasets. The MapReduce CPSO is the optimization and parallel methodology technique which is used to obtain the best clustering results. In MR-CPSO the clustering is made based on the centroid value. Since MR-CPSO is the optimization technique it is used to find the optimal centroids based on the velocity of the particle. The centroid value for each iteration is updated using particle's velocity. Since MR-CPSO is the parallel methodology it is used to reduce the time and space complexity. This MR-CPSO can be applied to both real-world and synthetic datasets. This MapReduce CPSO can work well with the increasing data sizes which is used to increase the cluster quality with minimal time and space requirement.

REFERENCES

[1] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is 'Nearest Neighbor' Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), pp. 217-235, 1999.

[2] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," Data Mining Knowledge Discovery, vol. 2, no. 4, pp. 311-324, 1998.

[3] K. Wang, S. Zhou, and J. Han, "Profit Mining: From Patterns to Actions," Proc. Eighth Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT), pp. 70-87, 2002.

[4] K. Wang, S. Zhou, Q. Yang, and J.M.S. Yeung, "Mining Customer Value: From Association Rules to Direct Marketing," Data Mining Knowledge Discovery, vol. 11, no. 1, pp. 57-79, 2005.

[5] H.-P. Kriegel et al., "Future Trends in Data Mining," Data Mining Knowledge Discovery, vol. 15, no. 1, pp. 87-97, 2007.

[6] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut, "Data Peeler: Constraint-Based Closed Pattern Mining in N-Ary Relations," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 37-48, 2008.

[7] K. Sim, Z. Aung, and V. Gopakrishnan, "Discovering Correlated Subspace Clusters in 3D Continuous-Valued Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 471-480. 2010.

[8] L. Ji, K.-L. Tan, and A.K.H. Tung, "Mining Frequent Closed Cubes in 3D Data Sets," Proc. 32nd Int'l Conf. Very Large Databases (VLDB), pp. 811-822, 2006.

[9] C.H. Cheng, A.W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 84-93, 1999.

[10] K. Sim, A.K. Poernomo, and V. Gopalkrishnan, "Mining Actionable Subspace Clusters in Sequential Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 442-453. 2010,

[11] B.J. Grant et al., "Novel Allosteric Sites on Ras for Lead Generation," PLoS One, vol. 6, no. 10, p. e25711, 2011.

[12] R. Gupta et al., "Quantitative Evaluation of Approximate Frequent Pattern Mining Algorithms," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 301-309, 2008.

[13] De Lathauwer L. (2000) "A Multilinear Singular Value Decomposition," SIAM J. Matrix Analysis Applications, vol. 21, no. 4, pp. 1253-1278.

[14] Fromont E., Prado A. and Robardet C. (2009.) "Constraint-Based Subspace Clustering," Proc. SIAM Int'l Conf. Data Mining (SDM),pp. 26-37.

[15] Kailing K., Kriegel H.P., Kroger P., and Wanka S. (2003) "Ranking Interesting Subspaces for Clustering High Dimensional Data,"Proc. Practice of Knowledge Discovery in Databases (PKDD), pp. 241- 252.

[16] Sequeira K. and Zaki M.J. (2004) "SCHISM: A New Approach for Interesting Subspace Mining," Proc. IEEE Fourth Int'l Conf. Data Mining (ICDM), pp. 186-193.

[17] Sun J., Tao D., and Faloutsos C. (2006) "Beyond Streams and Graphs: Dynamic Tensor Analysis," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 374-383.

[18] Wang K., Zhou S., and Han J. (2002) "Profit Mining: From Patterns to Actions," Proc. Eighth Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT), pp. 70-87.

