Asian Journal of Computer Science and Technology (AJCST)
Vol.2.No.1 2014 pp 49-52
available at: www.goniv.com
Paper Received :05-03-2014
Paper Published:28-03-2014
Paper Reviewed by: 1. John Arhter 2. Hendry Goyal
Editor : Prof. P.Muthukumar

AN EFFICIENT APPROACH FOR SUBSPACE CLUSTERING BY USING CAT SEEKER

M. Aravindan
Computer science & Engineering
Oxford Engineering College
Trichy

ABSTRACT

Identify several clustering problems, which require the mining of actionable subspaces defined by objects and attributes over a progression of time. These subspaces are actionable in the sense that they have the ability to propose profitable action for the decision-makers. We suggest mining actionable subspace clusters from sequential data, which are subspaces with high and associated utilities.In this proposed algorithm CAT Seeker, which uses a hybrid of SVD, optimization algorithm, and 3D frequent item set mining algorithm to mine CATSs in an efficient and parameter insensitive way.Propose to mine actionable subspace clusters from 3D dimensional data, which are subspaces with high and correlated utilities. The CAT Seeker has provide high efficiency in financial/ biological data and is able to discover to significant clusters and also find the fixed and optimal centroid values to cluster the data with efficient performance and reduceable size. Show that our clustering results are not sensitive to the framework parameters. In our case-study, Show that clusters with higher utilities correspond to higher action ability, and able to use our clusters to perform better than one of the most famous value investment strategies.

Index Terms—3D subspace clustering, singular vector decomposition, numerical optimization, protein structural and dynamics analysis, financial data mining

1. INTRODUCTION

CLUSTERING aims to find groups of similar objects and due to its usefulness, it is popular in a large variety of domains, such as geology, marketing, etc. Over the years, the increasingly effective data gathering has produced many high-dimensional data sets in these domains. As a consequence, the distance (difference) between any two objects becomes similar in high dimensional data, thus diluting the meaning of cluster]. A way to handle this issue is by clustering in subspaces of the data, so that objects in a group need only to be similar on a subset of attributes (subspace), instead of being similar across the entire set of attributes. The high-dimensional data sets in these domains also potentially change over time. We define such data sets as three-dimensional (3D) data sets, which can be generally expressed in the form of object-attribute-time, e.g., the stock-ratio-year data in the finance domain, and the residues-position-time protein structural data in the biology domain, among others.

2. PROBLEM MOTIVATION

The problems of usefulness and usability of subspace clusters are very important issues in subspace clustering. The usefulness of subspace clusters, and in general of any mined patterns, lies in their ability to suggest concrete actions. Such patterns are called actionable patterns, and they are normally associated with the amount of profits or benefits that their suggested actions bring. The usability of subspace clusters can be increased by allowing users to incorporate their domain knowledge in the clusters. To achieve usability, we allow users to select their preferred objects as centroids, and we cluster objects

that are similar to the centroids. In this paper, we identify real-world problems, which motivate the need to infuse subspace clustering with actionability and users' domain knowledge via centroids.

Financial example. Value investors scrutinize fundamentals or financial ratios of companies, in the belief that they are crucial indicators of their future stock price movements . For example, if investors know which particular financial ratio values will lead to rising stockprice, they can buy stocks having these values of financial ratio to generate profits. Experts like Graham have recommended certain financial ratios and their respective values. For example, Graham prefers stocks whose Price-Earnings ratio (measures the price of the stock in relative tothe earnings of the stock) is not more than 7. However, thereis no concrete evidence to prove their accuracy and the selection of the right financial ratios and their values has remained subjective. On the other hand, investors usually know a (limited) number of profitable stocks and these stocks can be used ascentroids to find other stocks that are fundamentally similar (in the context of financial ratios) to the centroids, and at thesame time, are profitable. Through this way, investors can understand which values of financial ratios are related to high price returns.

3. RELATED WORK

Majority of the subspace clustering algorithms handle 2D data , i.e., data having two dimensions, namely object and attribute. More recently, algorithms have been proposed to handle 3D data i.e., data having an additional context dimension (typically time or location). The solutions in mine subspace clusters in 3D binary data, thus they are not suitable for the more complicated 3D continuous-valued data. Xu et al. mine 3D subspace clusters that are non-axis-parallel, so it is not within our scope. Only algorithms GS-search , TRICLUSTER, MASC, and MIC mine subspace clusters in 3D continuous-valued data.

GS-search and MASC "flatten" the continuous valued 3D data set into a data set with a single time stamp.They require the clusters to occur in every time stamp, and it is hard to find clusters in data set that has a large number of time stamps. CATSeeker, TRICLUSTER, and MIC have the concept of subspace in all three dimensions i.e., they mine 3D subspace clusters that are subsets of attributes and subsets of time stamps.

TRICLUSTER is the pioneer work on mining 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 is a parameter that is difficult to set; at any distance thresholdsetting, 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.

Algorithm MIC proposed mining significant 3D subspace clusters in a parameter insensitive was a significant cluster are intrinsically prominent in the data, and they are usually small in numbers. There are also works that use the concept of significance, but they focus on mining interesting subspaces or significant subspaces , 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. Only CATSeeker and MASC can achieved these. However, CATSeeker is better than MASC, in the handling of subspace clusters in 3D data and in term of efficiency and scalability.

1. CATSeeker mines CATSs, which are subspace clusters in 3D subspaces, while MASC mines subspace clusters, which must occur in every time stamp of the data set.

2. CATSeeker uses a SVD-based algorithm to effectively prune the search space, while MASC does noprune the search space.

3. CATSeeker is guaranteed to be times faster than MASC, where is the number of attributes. For each centroid, MASC needs to run the optimization algorithm times, whereas CATSeeker only needs to run it once. There is constraint subspace clustering, and constraint is similar to actionability, as both dictate the clustering in a semi-supervised manner. However, constraints are indicators if objects should be clustered together, while utilities (that represent actionability) are continuous values indicating the quality of the objects. In summary, there lacks a centroids based, actionable3D subspace clustering algorithm that is

parameter insensitive and efficient. CATSeeker can effectively achieve all these.

4. LIMITATIONS

Existing 3D subspace clustering algorithms are Inadequate in mining actionable 3D subspace clusters.

Domain knowledge incorporation. In protein structural data, biologists need to know what residues potentially regulate the specified residue(s), and in stock data, investors want to find stocks which are similar in profit to the preferred stock of the investor. Hence, users' domain knowledge can increase the usability of the clusters. In addition, users should be allowed to select the utility function suited for the clustering problem.

3D subspace generation. In protein structural data, the residues do not always have the same dynamics across time. In stock data, stocks are homogeneous only in certain periods of time [12]. Hence, a true 3D subspace cluster should be in a subset of attributes and a subset of time stamps. Algorithm GS-search and MASC do not generate true 3D subspace clusters but 2D subspace clusters that occur in every time stamps.

Parameter insensitivity The algorithm should not rely on users to set the tuning parameters [6], or the results should be insensitive to the tuning parameters. Algorithm GS-search and Tricluster require users to tune parameters which strongly influence the results.

Actionable. Actionability, that was first proposed infrequent patterns and in subspace clusters, is the Ability to generate benefits/profits.

5. PROPOSED SYSTEM

This paper is a substantial extension of an earlier work ,and the differences are explained in. The following summarizes our contributions:

We identify the need to mine CATSs, which are clusters of objects that suggest profits or benefits to users, and users are allowed to incorporate their domain knowledge, by selecting their preferred objects as centroids of the clusters.

We propose algorithm CATSeeker, which uses a hybrid of SVD, optimization algorithm, and 3D frequent itemset mining algorithm to mine CATSs in an efficient and parameter insensitive way. We conduct a comprehensive list of experiments to verify the effectiveness of CATSeeker and to demonstrate its strengths over existing approaches:

- **Robustness.** Correct clusters are found using CATSeeker, even with 20 percent perturbation in the data.
- **Parameter insensitivity**. Correct clusters are found across diverse settings of CATSeeker's Tuning parameters.
- **Effectiveness.** CATSeeker has on average 180 percent higher accuracy in recovering Embedded clusters than current subspace clustering Algorithms.
- **Efficiency**. CATSeeker is at least 2 orders of Magnitude faster than the other centroids-based Subspace clustering algorithm MASC.
- **Applications on real-world data**. We show that CATSeeker has 82 percent higher profit/risk ratio than the next best approach in financial Data, and is able to discover biologically Significant clusters where other approaches have not succeeded.

6. CONCLUSION

Mining actionable 3D subspace clusters from continuousvalued 3D (object-attribute-time) data is useful in domains ranging from finance to biology. But this problem is nontrivial as it requires input of users' domain knowledge, clusters in 3D subspaces, and parameter insensitive and efficient algorithm. We developed a novel algorithmCATSeeker to mine CATS, which concurrently handles the multifacets of this problem. In our experiments, we verified the effectiveness of CATSeeker in synthetic and real world data. In protein application, we show that CATSeeker is able to discover biologically significant clusters (particularly residues that form potential drug binding site) while other approaches have not succeeded. In financial application, we show that CATSeeker is 82 percent better than the next best competitor in the return/risk (maximizing profits over risk) ratio. For future work, we plan to develop an algorithm where the optimal centroids are mined during the clustering process, instead of using fixed centroids.

REFERENCES

[1] Kelvin Sim, Ghim-Eng Yap, David R. Hardoon, Vivekanand Gopalkrishnan,,Gao Cong, and Suryani Lukman, "Centroid based actionable 3D subspace clustering by using CATseeker," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, June 2013

[2] H.-P. Kriegel, P. Kro¨ger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," ACM Trans.Knowledge Discovery from Data, vol. 3, no. 1, pp. 1-58, 2009.

[3] 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.

[4] 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.

[5] J.Y. Campbell and R.J. Shiller, "Valuation Ratios and the Long Run Stock Market Outlook: An Update," Advances in Behavioral Finance II, Princeton Univ. Press, 2005.

[6] 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,

[7] L. Zhao and M.J. Zaki, "TRICLUSTER: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 694-705. 2005,

[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.