ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Optimal Centroid Estimation Scheme for Multi Dimensional Clustering

K. Lalithambigai[1], Mr. S. Sivaraj, ME[2]

II-M.E (CSE), Dept. of CSE, SSM College of Engineering, Komarapalayam, Tamilnadu, India[1]

Assistant Professor, Dept. of CSE, SSM College of Engineering, Komarapalayam, Tamilnadu, India[2]

Abstract: High dimensional data values are processed and optimized with feature selection process. A feature selection algorithm is constructed with the consideration of efficiency and effectiveness factors. The efficiency concerns the time required to find a subset of features. The effectiveness is related to the quality of the subset of features.

3 dimensional data models are constructed with object, attribute and context information. Cluster quality is decided with domain knowledge and parameter setting requirements. CAT Seeker is a centroid-based actionable 3D subspace clustering framework. CAT Seeker framework is used to find profitable actions. Singular value decomposition, numerical optimization and 3D frequent itemset mining methods are integrated in CAT Seeker model. Singular value decomposition (SVD) is used to calculating and pruning the homogeneous tensor. Augmented Lagrangian Multiplier Method is used to calculating the probabilities of the values. 3D closed pattern mining is used to fetch Centroid-Based Actionable 3D Subspaces (CATS).

Optimal centroid estimation scheme is used to improve the financial data analysis process.. Intra cluster accuracy factor is used to fetch centroid values. Inter cluster distance is also considered in centroid estimation process. Dimensionality analysis is applied to improve the subspace selection process.

I. 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 astronomy, physics, geology, marketing, etc. Over the years, data gathering has become more effective and easier, resulting in many of these domains having high dimensional databases. As a consequence, the distance between *any* two objects becomes similar in high dimensional data, thus diluting the meaning of cluster. One way to handle this issue is by clustering in subspaces of the dimension space, so that objects in a group need only be similar on some subset of attributes, instead of being similar across the entire set of attributes.

Besides being high-dimensional, the databases in these domains also potentially change over time. In such sequential databases, finding subspace clusters per timestamp may produce a lot of spurious and arbitrary clusters, hence it is desirable to find clusters that persist in the database over some given period. Moreover, the usefulness of these clusters, and in general of any mined patterns, lies in their ability to suggest concrete and useful actions. Such patterns are called *actionable* patterns and they are normally associated with the amount of profit that their suggested actions bring. In this system identify real-world problems, particularly in the financial world, which motivates the need to infuse subspace clustering with action ability.

II. RELATED WORK

Majority of the subspace clustering algorithms handle 2D data [3], i.e., data having two dimensions, namely object and attribute. More recently, algorithms have been proposed to handle 3D data [7], i.e., data having an additional context dimension (typically time or location). The solutions in [4] mine subspace clusters in 3D binary data, thus they are not suitable for the more complicated 3D continuous-valued data. Xu et al. [5] 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, along with most of the subspace clustering algorithms, are parameter based 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 [1]. For example, the distance threshold 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.

Algorithm MIC proposed mining significant 3D subspace clusters in a parameter insensitive way. Significant clusters 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 terms of efficiency and scalability. 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 are continuous values indicating the quality of the objects. In summary, there lacks a centroid based, actionable 3D subspace clustering algorithm that is parameter insensitive and efficient. CATSeeker can effectively achieve all these.

III. 3D SUBSPACE CLUSTERING

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 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, instead of being similar across the entire set of attributes [2]. 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. In such data sets, finding subspace clusters per time stamp may produce a lot of spurious and arbitrary clusters, hence it is desirable to find clusters that persist in the database over a given period.

The problems of usefulness and usability of subspace clusters are very important issues in subspace clustering [2]. 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. The usability of subspace clusters can be increased by allowing users to incorporate their domain knowledge in the clusters [6]. 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. 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 stock price, 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 is not more than 7. However, there is no concrete evidence to prove their accuracy, and the selection of the right financial ratios and their values has remained subjective.

Biologists are interested in finding regulating residues that can regulate catalytic residue(s) and these regulating residues have the following two properties. They are Actionable and Homogeneous. Flexibility and dynamics are properties of biological molecules, e.g., proteins. The flexibility of the residues are indicated by their B-factor, and the dynamics of the residues are indicated by their positional dynamics across time. The catalytic residues can be used as centroids, to find regulating residues that have similar dynamics with the centroids and are as flexible as their centroids. These two examples highlight the needs to find actionable clusters of objects that suggest profits or benefits and to substantiate their actionability, these clusters should be homogeneous and correlated across time. In addition, users should be allowed to incorporate their domain knowledge, by selecting their preferred objects as centroids of the actionable 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 [6]. 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. 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 in frequent patterns and in subspace clusters is the ability to generate benefits/profits.

We propose mining Centroid-based, Actionable 3D Subspace clusters with respect to a set of centroids, to solve the above issues. CATS allows incorporation of users' domain knowledge, as it allows users to select their preferred objects as centroids, and preferred utility function to measure the actionability of the clusters. 3D subspace generation is allowed, as CATS is in subsets of all three dimensions of the data. Mining CATSs from continuous-valued 3D data is nontrivial, and it is necessary to breakdown this complex problem into subproblems: 1) pruning of the search space, 2) finding subspaces where the objects are homogeneous and have high and correlated utilities, with respect to the centroids, and 3) mining CATSs from these subspaces. We propose a novel algorithm, CATSeeker, to mine CATSs via solving the three subproblems:

- CATSeeker uses SVD to prune the search space, which can efficiently prune the uninteresting regions, and this approach is parameter free.
- CATSeeker uses augmented Lagrangian multiplier method to score the objects in subspaces where they are homogeneous and have high and correlated utilities, with respect to the centroids. This approach is shown to be parameter insensitive.
- CATSeeker uses the state of the art 3D frequent itemset mining algorithm to efficiently mine CATSs, based on the score of the objects in the subspaces.

IV. PROBLEM STATEMENT

Object, attribute and context information are linked in the 3 dimensional data models. Cluster quality is decided with domain knowledge and parameter setting requirements. CAT Seeker is a centroid-based actionable 3D subspace clustering framework. CAT Seeker framework is used to find profitable actions. Singular value decomposition, numerical optimization and 3D frequent itemset mining methods are integrated in CAT Seeker model. Singular value decomposition (SVD) is used to calculating and pruning the homogeneous tensor. Augmented Lagrangian Multiplier Method is used to calculating the probabilities of the values. 3D closed pattern mining is used to fetch Centroid-Based Actionable 3D Subspaces (CATS). The following problems are identified in the CAT Seekar model. They are fixed centroid model, limited cluster accuracy, inter cluster distance is not focused and dimensionality is not optimized.

V. MULTI DIMENSIONAL CLUSTERING WITH OPTIMAL CENTROIDS

The proposed system is designed to analyze the stock market data values. CAT Seeker is improved with optimal centroid values. Profitable actions are identified from the cluster results. The system is divided into five major modules. They are cube construction process, clustering with fixed centroid, optimal centroid estimation, clustering with dynamic centroid and action identification. Cube construction process is applied to collect 3D data values. Fixed centroid based clustering approach is used to partition the data values. Optimal centroid selection process is designed with cluster distance factors. Dynamic centroid based clustering is performed with optimal centroid values. Pattern mining is used to identify the profitable actions.

5.1. Cube Construction Process

The data cube is constructed using the stock market transaction details. Share price details are collected from the National Stock Exchange (NSE) and Bombay Stock Exchange (BSE). Opening price, closing price, high price and low price levels are collected for a set of companies. Data cube is formed for a set of companies to a period of time. We deal with a continuous valued, 3D data D, with its dimensions defined by objects O, attributes A, and time stamps T. Let the value of object o on attribute a hand in time stamp t be denoted v_{oat}. We denote feature (a, t) as a pair of attribute a and time stamp t. Let c be an object selected as the centroid. We denote $h_c(v_{oat})$ as a homogeneous function to measure the homogeneity between object o and centroid c, on attribute a and in time stamp t. The gist of this algorithm is in using the variance of the homogeneity values to guide the pruning process. By using SVD on the matrix M, we can calculate the variance of the homogeneity values of each row or column of M. A row or column that contains high homogeneity values has high variance, as its values are away from the dummy "0" values. Therefore, we keep those rows or columns that have high variance, and discard the rest. In the homogeneous matrix M, we keep rows o_1, o_2, o_3 and column (a_1, t_1) as they have high variances, and prune the rest. Instead of matrix SVD, we could use tensor SVD, which does not unfold the tensor to matrix. However, tensor SVD is too aggressive in its pruning, as removing an object, attribute or time means removing a matrix of the tensor.

5.2. Clustering With Fixed Centroid

Clustering process is applied on the financial data cube. Cluster centroids are randomly initialized for each cluster. CATSeeker algorithm is used for the clustering process. Singular value decomposition (SVD) pruning and Bound-Constrained Lagrangian Method (BCLM) algorithms are used in the pruning and probability estimation process. Calculating and pruning the homogeneous tensor using SVD. Given a centroid c, we define a homogeneous tensor S ε $[0, 1]^{|O|\times|A|\times|T|}$, which contains the homogeneity values s_{oat} with respect to centroid c. The first data set of a 3D continuous-valued data set with centroid o_5, and the second data set shows its homogeneous tensor. Mining CATSs from the high-dimensional and continuous- valued tensor S is a difficult and time-consuming process. Hence, it is vital to first remove regions that do not contain CATSs. A simple solution is by removing values soat that are less than a threshold, but it is impossible to know the right threshold. Hence, we propose to efficiently prune tensor S in a parameter-free way, by using the variance of the data to identify regions of high homogeneity values soat. The constraint function g(P) is a summation of probabilities, and it is possible that only the probabilities involving the centroid are nonzeros and the rest are zeros. One remedy is to use a multiplication of probabilities for the constraint function, to ensure all probabilities are nonzeros. However, we do not force clusters to be created, as it is possible that a centroid is highly dissimilar to other objects. The clusters can be mined in both synthetic and real world data sets, which mean that our approach does not give trivial solutions. We use the augmented Lagrangian multiplier method, known as the Bound-Constrained Lagrangian Method (BCLM), to optimize F(P). BCLM exploits the smoothness of both f(P) and g(P) and replaces our constrained optimization problem with iterations of unconstrained optimization subproblems, and the iterations continue until the solution converges.

5.3. Optimal Centroid Estimation

The optimal centroid estimation scheme is used to initialize the centroid values for the clusters. Centroid estimation process is enhanced with distance analysis mechanism. Intra cluster and inter cluster relationships are analyzed in the centroid estimation process. Transaction relationship is also considered in the centroid estimation process. We denote homogeneous value soat as the output of the homogeneous function hc(v_{oat}), i.e., hc(v_{oat}) = s_{oat}. We allow users to define the homogeneous function, but the homogeneous values must be normalized to [0, 1], such that s_{oat} = 1 indicates that the value voat is “perfectly” homogeneous to the value of the centroid vcat, while s_{oat} = 0 indicates otherwise. We use the Gaussian function as the homogeneous function, as it normalizes the similarity between object o and centroid c on feature (a, t), to [0, 1]. We also randomly selected 10 percent of the objects as centroids in each data set, and evaluated the quality of the clusters mined using them. We developed a novel algorithm CATSeeker 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 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 ratio.

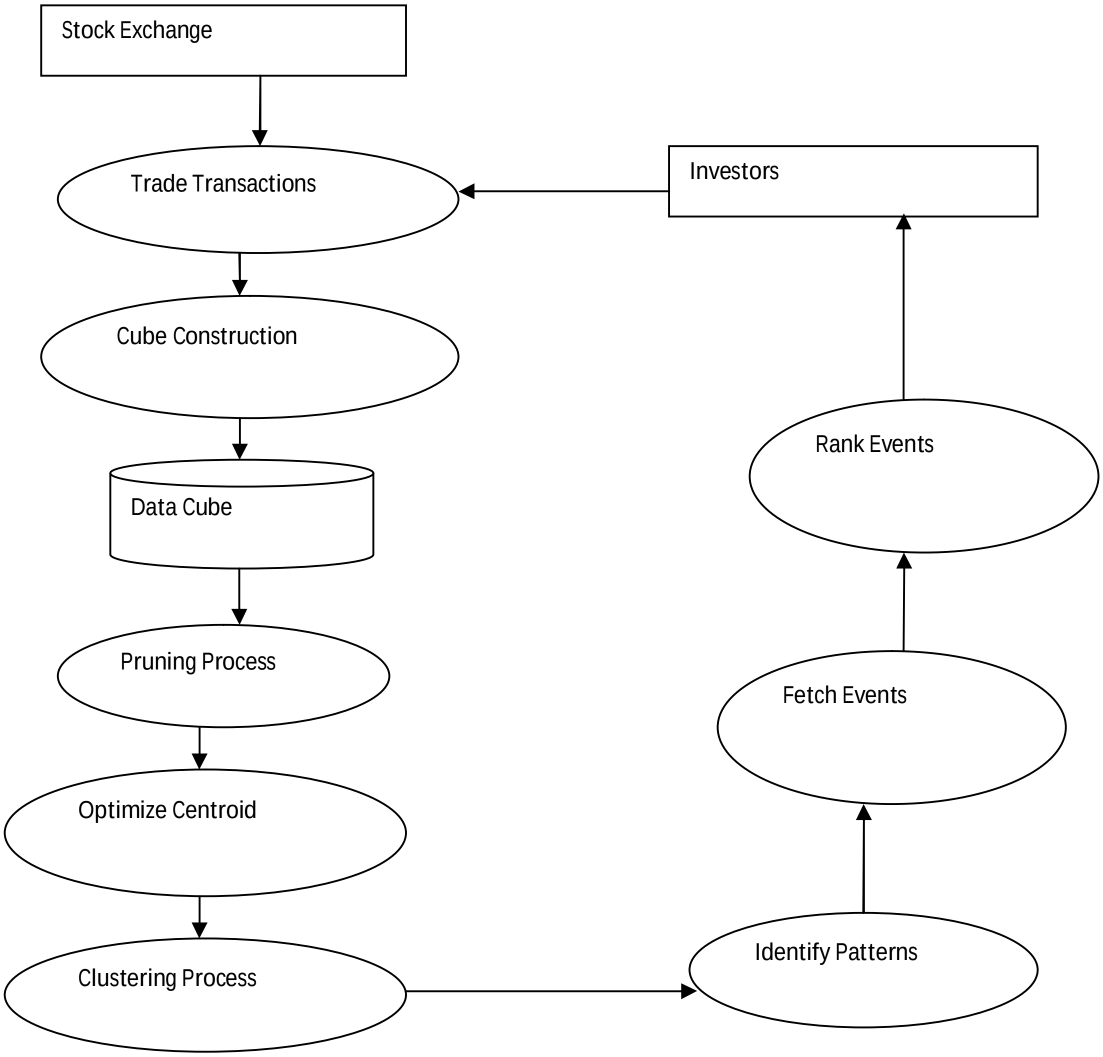

Fig. No: 5.1. Multi Dimensional Clustering with Optimal Centroids

5.4. Clustering With Dynamic Centroid

Three dimensional data clustering is performed on subspaces. Distance based centroid model is used in the clustering process. Centroid optimization process is performed in all cluster iterations. Fitness functions are used to verify the data assignment process. Calculating the probabilities of the values using the augmented Lagrangian Multiplier Method. We use the homogeneous tensor S with the utilities of the objects to calculate the probability of each value voat of the data to be clustered with the centroid c. We map this problem to an objective function, and use the augmented Lagrangian Multiplier Method to maximize this function. This approach is robust to perturbations in data and less sensitive to the input parameters. We created synthetic data sets with embedded clusters, and used the embedded clusters as the "ground truth" to evaluate the quality of the clusters mined by the different algorithms. We also studied the effectiveness of the SVDpruning of CATSeeker by comparing it with 1) CATSeeker without SVDpruning and 2) CATSeeker with simple pruning. In CATSeeker with simple pruning, values below a threshold in the homogeneous tensor are pruned. TRICLUSTER and MaxnCluster have seven and three parameters, respectively, and it is hard to enumerate all possible settings.

5.5. Action Identification

Profitable actions are identified from the clustered data values. Transaction patterns are used in the action identification process. 3 Dimensional Closed Frequent Itemset (3D CFTI) mining algorithm is used for the action detection process. Actions are listed with reference to the profit ratio levels. Mining CATSs using 3D closed pattern mining. After calculating the probabilities of the values, we binarize the values that have high probabilities to "1". We then use efficient 3D closed pattern mining algorithms efficiently mine subcuboids of "1", which correspond to the CATSs. An example of CATS is last data set. A compound event is a set of primary events taken from different random variables. More precisely, it is a realization of X^s and is denoted by x^s. The order of the compound event is |s|. Such a difference is minor, since one can always map all primary events to items by considering each primary event as an attribute-value pair. Since the proposed method is designed to cluster the patterns produced by PD, the notations of PD and are adopted. PD uncovers compound events that do not follow a preassumed model. Any default model can be chosen according to the problem domain and the available knowledge. If a priori knowledge about the domain is not available, similar to chi-square statistic, a model assuming the independence of the random variables is normally used.

VI. CONCLUSION

Three subspace clustering techniques are used to partition the transactions with action identification process. CAT Seeker framework is used to fetch Centroid Actionable 3D Subspace clusters. Optimal centroid estimation scheme is integrated with CAT Seeker framework. Cluster accuracy is improved with efficient inter cluster distance model. High feature selection quality is achieved by the system. Process time is low in the optimal centroid based scheme. High cluster accuracy is achieved by the system. Inter cluster distance is optimized by the dynamic centroid selection scheme.

REFERENCES

[1] J. Nocedal and S.J. Wright, Numerical Optimization, pp. 497-528. Springer, 2006.

[2] H.-P. Kriegel, P. Kroger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," ACM Trans. Knowledge Discovery from Data, , 2009.

[3] G. Moise and J. Sander, "Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2008.

[4] L. Cerf, and J.-F. Boulicaut, "Data Peeler: Constraint-Based Closed Pattern Mining in N-Ary Relations," Int'l Conf. Data Mining, 2008.

[5] X. Xu, Y. Lu, K.-L. Tan, and A.K.H. Tung, "Finding Time-Lagged 3D Clusters," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009.

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

[7] and Scholkopf, "Multi-Way Set Enumeration in Weight Tensors," Machine Learning, 2010.