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

Integration of Subspace Clustering and Action Detection on Financial Data

Meenu Mathai[1], Mrs.P.Sumathi[2]

Student IInd year M.E , Department of CSE, KSRIET,Tiruchengode, Tamilnadu, India[1]

Assistant Professor, KSR Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India[2]

Abstract—Object, attribute and context information are linked in the dimensional data models. Cluster quality is decided with domain knowledge and parameter setting requirements. CAT Seeker is a centroid-based actionable D subspace clustering framework. CAT Seeker framework is used to find profitable actions. Singular value decomposition, numerical optimization and D frequent itemset mining methods are integrated in CAT Seeker model. CAT Seeker framework is improved with optimal centroid estimation scheme. 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. Experimental results on financial data show that CATSeeker with optimal centroid significantly outperforms all the competing methods in terms of efficiency, parameter insensitivity, and cluster usefulness.

IndexTerms—Dsubspaceclustering,singularvectordecomposition,numericaloptimization and financialdatamining

I . INTRODUCTION

CLUSTERINGaimstofindgroupsofsimilar objectsanddue to its usefulness, it is popular in a large variety of domains,such asgeology, marketing,etc.Over theyears, theincreasinglyeffectivedatagatheringhasproducedmany high-dimensionaldata sets in these domains.As acon-sequence,thedistance(difference)betweenanytwoobjects becomes similar inhighdimensional data, thus dilutingthe meaningof cluster.A way to handlethis issue is by clusteringinsubspacesofthedata,sothatobjectsinagroup need onlytobesimilar onasubset ofattributes(subspace), insteadofbeing similar across the entire set ofattributes(fullspace)[1].

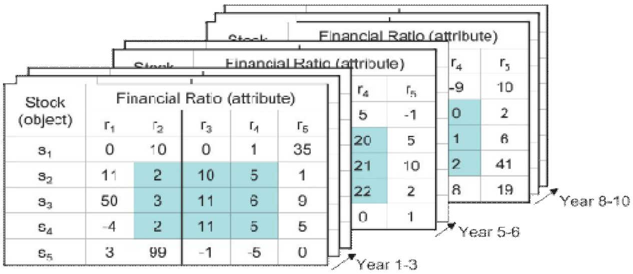
Thehigh-dimensional data setsinthese domainsalso potentiallychange over time. Wedefine such data sets as three-dimensional (D)data sets,which canbegenerally expressed in the form of object-attribute-time, e.g., the stock-ratio-yeardata in the finance domain,and the residues-position-timeproteinstructuraldata in the biology domain,amongothers.

In such data sets, finding subspace clusters per time stampmayproducealotofspuriousand arbitrary clusters, hence it is desirable to find clusters that persist in the databaseover agiven period.

Theproblemsofusefulnessand usabilityofsubspaceclusters areveryimportant issues insubspaceclustering .The usefulnessofsubspaceclusters, andingeneralofanymined patterns,lies in their ability to suggestconcrete actions. Suchpatternsarecalled actionablepatterns,and they are normallyassociatedwith the amountofprofits orbenefits that their suggestedactions bring .

In thispaper,we identifyreal-worldproblems,whichmotivatetheneed toinfuse subspaceclustering with actionabilityand users' domainknowledgeviacentroids.Financial example. Value investorsscrutinizefundamentalsorfinancial ratiosofcompanies,inthe belief that theyarecrucial

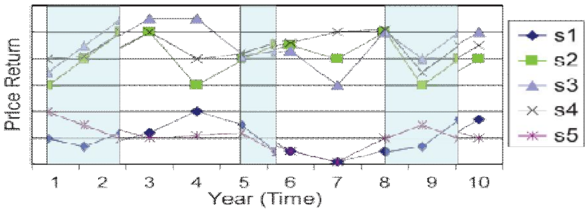
indicatorsoftheir futurestockprice movements,.Forexample,ifinvestorsknow which particularfinancial ratio values will lead to rising stock price, they canbuy stocks havingthese values offinancial ratiotogenerateprofits. Experts likeGraham have recommended certain financial ratios and their respective values. Forexample, Grahamprefersstocks whosePrice- Earningsratio (measuresthepriceofthestockinrelative to theearningsofthestock)isnotmorethan.However,there isnoconcrete evidence toprove theiraccuracy,and the selection ofthe right financial ratios and their values has remainedsubjective.



Stock (object)	Financial Ratio (attribute)				
	r_1	r_2	r_3	r_4	r_5
s_1	0	10	0	1	35
s_2	11	2	10	5	1
s_3	50	3	11	6	9
s_4	-4	2	11	5	5
s_5	3	99	-1	-5	0

Year 1-3
Financial Ratio (attribute)
r_4 r_5
5 -1
20 5
21 10
22 2
0 1
Year 5-6
Financial Ratio (attribute)
r_4 r_5
-9 10
0 2
1 6
2 41
8 19
Year 8-10

(a)



(b)

Fig. 1a shows a 3D stock-financial ratio-year financial data.Let us assume that an investor uses s2 (that is worthinvesting, e.g., Apple) as a centroid. The shaded regionshows a cluster of stocks with centroid s2, and they arehomogeneous in the subspace defined by financial ratiosr2; r3; r4 and year 1-3, 5,6, 8-10. This cluster of stocks areactionable (can generate profits) as shown by their high andcorrelated price returns ((sold price of stock – purchasedprice of stock)/purchased price of stock) (see Fig. 1b).

II . EXISTING SYSTEM

We denote clusters as centroid-based, actionable 3D subspace clusters (CATSs), and we also denote utility as a function measuring the profits or benefits of the objects. A CAT should have the following properties:

1. Its objects have high and correlated utilities in a set of time stamps T, so that the action suggested by the cluster is profitable or beneficial to users.
2. Its objects exhibit a significant degree of homogeneity in the subspace defined by a set of attributes A, across the set of time stamps T. This ensures that the high utilities of its objects do not cooccur by chance.

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. 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, 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.

The 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, CAT Seeker, to mine CATSs via solving the three sub problems:

1. CATSeeker uses SVD to prune the search space, which can efficiently prune the uninteresting regions, and this approach is parameter free.
2. 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.
3. 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.

2.1 Drawbacks

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 drawbacks are identified in the existing system.

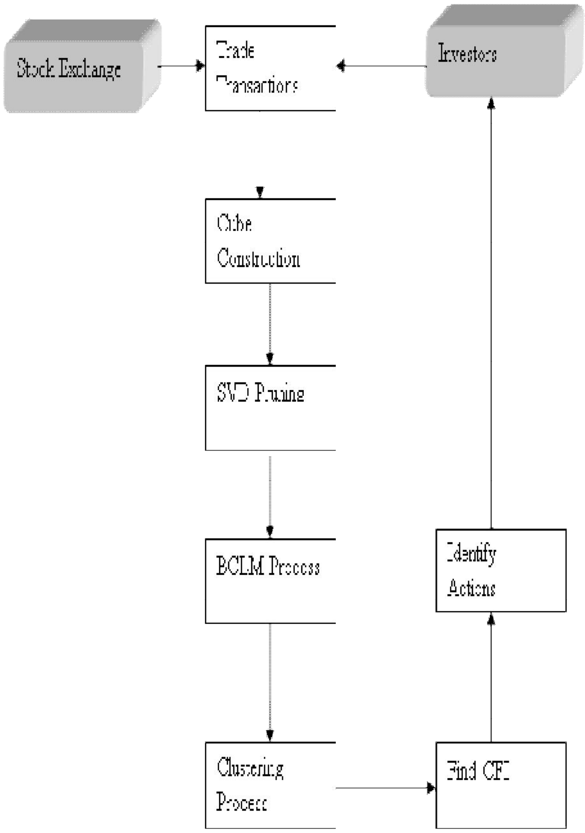
- Fixed centroid model
- Limited cluster accuracy
- Inter cluster distance is not focused
- Dimensionality is not optimized

III . PROPOSED SYSTEM

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. CAT Seeker framework is improved with optimal centroid estimation scheme. 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.

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.

Definition (CATS MINING PROBLEM). Given a contin- uous-valued 3D data set D and a set of centroids {c1 ; ... ; cn }, we set out to find all CATSs O x A x T with respect to the optimal centroids[1].

3.1 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.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

ALGORITHM CATSEEKER

The framework of CATSeeker is illustrated below, which consists of three main modules[1]:

1. Calculating and pruning the homogeneous tensorusing SVD. Given a centroid c, we define a homogeneous tensor which contains the homogeneity values soat with respect to centroid c. The first data set of shows a 3D continuous-valued data set with centroid 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.

2. Calculating the probabilities of the values using theaugmented 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 .

3. Mining CATSs using 3D closed pattern mining. Aftercalculating the probabilities of the values, we binarize the values that have high probabilities to "1"We then use efficient 3D closed pattern mining algorithms to efficiently minesubcuboids of "1", which correspond to the CATSs.

IV. CONLUSION

Mining actionable 3D subspace clusters from continuous valued 3D (object-attribute-time) data is useful in domain finance. 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 algorithm CATSeeker with optimal centroid to mine CATS, which concurrently handles the multifacets of this problem.In financial application, we show that CATSeeker is better than the next best competitor in the return/risk (maximizing profits over risk) ratio.

REFERENCES

[1] Kelvin Sim, and SuryaniLukman (2013), 'Centroid-Based Actionable 3D Subspace Clustering', IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no. 6.

[2] 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.

[3] 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.

[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] 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.