ISSN:0975-9646

An Efficient Actionable 3D Subspace Clustering Based on Optimal Centroids

V.Atchaya, C.Prakash
PG Scholar, CSE Department
Jerusalem College of Engineering, Chennai, India

Abstract- **An efficient Actionable 3D Subspace Clustering based on Optimal Centroids from continuous valued data represented three dimensionally which is suitable for real world problems profitable stocks discovery , biologically significant protein residues etc. It achieves actionable patterns ,incorporation of domain knowledge which allows users to choose the preferred utility(profit/benefit) function, parameter insensitivity, real world applications and excellent performance through a set of optimal centroids and by the combination of singular value decomposition, augmented lagrangian multiplier and 3D closed frequent item set mining.**

***Keywords*-actionable subspace clustering, financial mining,centroid based**

I. INTRODUCTION

Data mining is the process of extracting potentially useful information from a data set. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)[2]. When the dimensionality increases, usually only a small number of dimensions are relevant to certain clusters, but data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. Moreover, when dimensionality increases, data usually become increasingly sparse because the data points are likely located in different dimensional subspaces. When the data become really sparse, data points located at different dimensions can be considered as all equally distanced, and the distance measure, which is essential for cluster analysis, becomes meaningless.

Feature selection techniques are commonly utilized as a pre-processing stage for clustering, in order to overcome the curse of dimensionality. The most informative dimensions are selected by eliminating irrelevant and redundant ones. Such techniques speed up clustering algorithms and improve their performance. Nevertheless, in some applications, different clusters may exist in different subspaces spanned by different dimensions. In such cases, dimension reduction using a conventional feature selection technique may lead to substantial information loss

A way to handle this issue is by clustering in subspaces of the data[3], 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(full space) Subspace clustering is an extension to feature subset selection that has shown its strength at high-dimensional clustering. It is based on the observation that different subspaces may contain different, meaningful clusters. Subspace clustering searches for groups of clusters within different subspaces of the same data set.

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, users are allowed to select their preferred utility function and the centroids are calculated with respect to it, and we cluster objects that are similar to the centroids.

The high-dimensional data sets in these domains also potentially change over time. Such data sets are 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,

Our Contributions:

- We find out the need to mine actionable 3D subspaces with respect to a set of centroids generated based on the user’s domain knowledge of utility function.

- An algorithm is proposed which does:
  - A Homogeneous Tensor is formed with respect to the set of centroids generated.
  - Tensor is pruned using SVD.
  - Probabilities are calculated and goodness of clusters estimated
  - Binary transformation is applied and closed frequent pattern Mining applied to mine CATSs.
  - Significantly outperforms all the competing methods in terms of efficiency, parameter insensitivity, and cluster usefulness.

II.RELATED WORK

Most of the subspace clustering algorithms work well with data represented in two dimensional space and most of them not actionable. Traditional subspace clustering algorithms generate subspace clusters and to measure the goodness of the subspace clusters user has to set thresholds to measure but optimal thresholds is always a question mark.

The CLIQUE[3] algorithm was one of the first subspace clustering algorithms. The algorithm combines density and grid based clustering and uses an APRIORI style search technique to find dense subspaces.

PROjected CLUStering(PROCLUS)[3] was the first top-down subspace clustering algorithm. The distance based approach of PROCLUS is biased toward clusters that are hyper-spherical in shape. Also, while clusters may be found in different sub- spaces, the subspaces must be of similar sizes since the user must input the average number of dimensions for the clusters. Clusters are represented as sets of instances with associated medoids and subspaces and form non-overlapping partitions of the dataset with possible outliers. PROCLUS is actually somewhat faster than CLIQUE due to the sampling of large datasets. However, using a small number of representative points can cause PROCLUS to miss some clusters entirely.

Automatic subspace clustering of high dimensional data[16] the work in 1998 is to find clusters embedded in subspaces of high dimensional data without requiring the user to guess subspaces that might have interesting clusters. It didn't guarantee the quality of cluster.

Entropy-based subspace clustering for mining numerical data(1999)[17] to mine for interesting/significant subspaces and for measuring the goodness of clusters using the measure of entropy.

Ranking Interesting subspaces for clustering High dimensional Data(2003) to present a pre-processing step which detects all interesting subspaces of high dimensional data containing clusters(dense regions of arbitrary shape and size).

Density connected Subspace clustering for high dimensional data(2004)[18] to present the effective and efficient approach to subspace clustering problem using the concept of density-connectivity and is able to detect arbitrarily shaped and positioned clusters in subspaces.

A Fast algorithm for subspace clustering by pattern Similarity(2004)[19] To develop a pattern-based clustering methods i) to handle large datasets and ii) to discover pattern similarity embedded in
data sequences. Not feasible for data sets in high dimensional sets. Can't be applied to sequential data sets.

Efficient mining of Distance based subspace clusters(2009)[10] to use a sliding window approach to partition the dimensions to preserve significant clusters. Presence of highly overlapped clusters.

Bayesian Overlapping Subspace Clustering(2009)[11] to present a hierarchical generative model for matrices with potentially overlapping uniform sub-block structures. Computationally slow for large data sets.

K-Subspace Clustering(2009) to extend the K-means clustering algorithm to accommodate subspace clusters in addition to the usual ball-shaped clusters(line-shaped,plane-shaped,ball-shaped).

Discovering correlated subspace Clusters in 3D continuous valued data(2010)[5] to mine significant 3D subspace clusters in a parameter insensitive way used in microarray analysis and stock analysis.

Mining actionable subspace clusters in sequential data(2010)[6] to mine actionable(ability to suggest profitable action) subspace clusters defined by objects and attributes over a sequence of time. It flatten the continuous valued 3D data set with single timestamp so they are not efficient in generating 3D subspaces.

A feature group weighting method for subspace clustering of high-dimensional data (2011) To weigh subspaces in feature groups(the *features of high-dimensional data are divided into feature groups, based on their natural characteristics)* and individual features for clustering high-dimensional data.

Centroid based actionable 3D subspace clustering (2013)[1] to mine an actionable 3D subspace clusters from continuous valued 3D data with respect to the set of centroids suitable for real world applications. Use of Fixed centroid is the drawback.

III. PROPOSED SYSTEM

An Optimal Centroid based Actionable 3D subspace cluster framework consist of five main modules:

Algorithm OCATSEEKER:

1.Centroid Selection: Based on the user input a set of centroids are selected using Single pass seed selection algorithm.[8]

2.Similarity Tensor Formation: Based on the set of centroids a similarity tensor(multidimensional matrix) is formed using Gaussian function.

3.Pruning the tensor: The Homogeneous tensor formed is pruned using Singular value decomposition thereby the more similar values are retained and others are pruned with respect to the variance.

4.Calculating the probabilities of the Values using the Augmented
Lagrangian Multiplier Method: Using the homogeneous tensor with the utilities of the objects probabilities of each value of data to be clustered with the centroid is calculated. To measure the goodness Objective function is used and to maximize this function we use augmented Lagrangian Multiplier method[15].

5.Mining OCATSs: Binary Transformation is performed and then an efficient 3D closed frequent pattern mining[20] is used to mine the sub-cuboids which correspond to the Optimal Centroid based Actionable Three dimensional Subspace clusters(OCATSs).

Centroid Selection:

Algorithm1:Single Pass Seed Selection(SPSS)

- Selects the highest density point as the first centroid and also calculates the minimum distance automatically using highest density point which is close to more number of other points in the data set.

Algorithm 1 Single Pass Seed Selection

Input:
Financial Database
Output:
Set of centroids
Description:

```
1:  Calculate distance matrix $Dist_{mxm}$ in which $dist(X_i,X_j)$ represents distance from $X_i$ to $X_j$;
2:  Find Sumv in which Sumv(i) is the sum of the distances from $X_i$th point to all other points
3:  Find the index,h of minimum value of Sumv and find highest density point $X_h$;
4:  Add $X_h$ to C as the first centroid;
5:  For each point $X_i$, set d $(X_i)$ to be the distance between $X_i$ and the nearest point in C;
6:  Find y as the sum of distances of first m/k nearest points from the Xh;
7:  Find the unique integr I so that
8:  $d(X_1)^2+d(X_2)^2+...+d(X_i)^2> =y>d(X_1)^2+d(X_2)^2+...+ d(X_{(i-1)})^2$;
9:  Add $X_i$ to C;
10: Repeat steps 5-8 until k centroids are found.
```

Similarity Tensor Formation:[13]

Gaussian Function:To measure the similarity

$$h_c(v_{oat}) = exp\left(-\frac{|v_{cat} - v_{oat}|}{2\sigma_c^2}\right)$$

$v_{cat} \longrightarrow$ value of the centroid
$v_{oat} \longrightarrow$ value of the object-attribute-time

c is a parameter which controls the width of the Gaussian function, centred at centroid c.

$$\sigma_c = \frac{1}{k} \sum_{n \in Neigh_{cat}} dist_a(c, n),$$

Normally k=10.

Algorithm2:SVDpruning

Input:
Homogeneous tensor S
Output:
Pruned Homogeneous Tensor
Description:

```
1:  Unfold the homogeneous tensor into matrix using unfold function
2:  Add a dummy row and column to matrix to stretch the variance of values
3:  Perform Zero Mean Normalization
4:  By SVD[14] calculate the variance of homogeneity values
5:  Keep the rows and columns that have high variance and discard the rest
6:  After pruning fold back to homogeneous tensor using fold function
```

Calculating the Probabilities Of values:

Algorithm 3:

Input:
Pruned Homogeneous Tensor
Output:
Optimal probability distribution
Description:

```
1:  Calculate the probability of object to be clustered with centroid.
2:  Calculate the objective function to calculate our probabilities.
3:  Maximize the objective function using augmented Lagrangian multiplier method
4:  By iterating the above procedure optimal probability distribution is obtained.
```

Mining OCATSs:

Algorithm4:

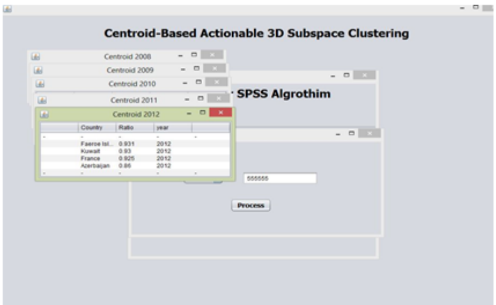
```
1:  Values with probabilities greater than initial probability are clustered with the centroid
2:  Binarize it by assigning 1 to the values greater than initial probability and the rest as 0
3:  Use 3D closed frequent item-set mining algorithm to mine sub-cuboids which correspond to CATs.
```

3D Closed Frequent Pattern Mining: Cube Miner:
Assign a Cutter '0' and the values with 0 are cut and the values 1 are retained and the final OCATSs are mined.

IV.EXPERIMENTS & RESULTS

The performance of OCATSeeker is tested using real financial data which shows the country ,region ,profit, income, ratio and year .Data set can be downloaded from http://econ.worldbank.org/ .Experiment is done using Netbeans IDE(java coding) and MYSQL server. Centroid Selection and Tensor Formation was tested with this data set .Many centroids are selected and the values that are similar (Similarity tensor) is also formed with respect to the centroids. The quality of the algorithm is compared with the CATSeeker[1] algorithm which takes centroid as user's domain knowledge incorporation and tensor is formed with respect to user's centroid selection. Here optimal centroids are generated using single pass seed selection algorithm in which its main objective is to select optimal centroids. SPSS algorithm's quality is already assessed with the traditional centroid based clustering algorithms and it proved to be the best.

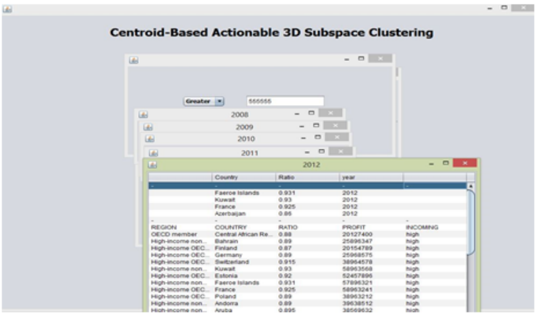
Chosen Centroids



	Country	Ratio	year	
-	-	-	-	-
	Faeroe Isl...	0.931	2012	
	Kuwait	0.93	2012	
	France	0.925	2012	
	Azerbaijan	0.86	2012	
-	-	-	-	-

Figure 1:Centroids generated

Homogeneous Tensor



	Country	Ratio	year	
-	-	-	-	-
	Faeroe Islands	0.931	2012	
	Kuwait	0.93	2012	
	France	0.925	2012	
	Azerbaijan	0.86	2012	
-	-	-	-	-
REGION	COUNTRY	RATIO	PROFIT	INCOMING
OECD member	Central African Re...	0.88	20127400	high
High-income non...	Bahrain	0.89	25896347	high
High-income OEC...	Finland	0.87	20154789	high
High-income OEC...	Germany	0.89	25968575	high
High-income OEC...	Switzerland	0.915	38964578	high
High-income non...	Kuwait	0.93	58963568	high
High-income OEC...	Estonia	0.92	52457896	high
High-income non...	Faeroe Islands	0.931	57896321	high
High-income OEC...	France	0.925	58963241	high
High-income OEC...	Poland	0.89	38963212	high
High-income non...	Andorra	0.89	39638512	high
High-income non...	Aruba	0.895	38569632	high

Figure 2:Homogeneous Tensor formation

V.CONCLUSION

Financial Data Mining of Actionable 3D subspace clusters based on Optimal Centroid from financial database is useful in domains ranging from finance to biology.We developed an algorithm OCATSeeker which first optimally selects set of centroids and form similarity tensor and uses techniques like Singular Value Decomposition,Augmented Lagrangian Multiplier,Binary Transformation and Cube Miner and efficiently mines 3D subspaces comparing to the other algorithms.

REFERENCES:

[1] K.Sim G.Yap "Centroid based Actionable 3D subspace clustering" IEEEE transactions on Knowledge and data engineering 2013. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] http://en.wikipedia.org/wiki/Cluster_analysis

[3] Jiawei Han, Micheline Kamber," Data Mining concepts and Techniques".

[4] D. Jiang, J. Pei, M. Ramanathan, C. Tang, and A. Zhang, "Mining Coherent Gene Clusters from Gene- Sample-Time Microarray Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 430-439. 2004.

[5] 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.

[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] Robust seed selection algorithm for k-means type algorithms International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 5, Oct 2011 K. Karteeka Pavan, 2Allam Appa Rao, 3A.V. Dattatreya Rao, G.R.Sridhar.

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

[10] Q. Fu and A. Banerjee, "Bayesian Overlapping Subspace Clustering,"Proc. IEEE Ninth Int'l Conf. Data Mining (ICDM), pp. 776-781,2009.

[11] G. Liu, K. Sim, J. Li, and L. Wong, "Efficient Mining of Distance-Based Subspace Clusters," Statistical Analysis Data Mining, vol. 2, nos. 5/6, pp. 427-444, 2009

[12] 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.

[13] E. Georgii, K. Tsuda, and B. Scho¨lkopf, "Multi- Way Set Enumeration in Weight Tensors," Machine Learning, vol. 82, pp. 123-155, 2010.

[14] L. De Lathauwer et al., "A Multilinear Singular Value Decomposition,"SIAM J. Matrix Analysis Applications, vol. 21, no. 4, pp. 1253-1278, 2000740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

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

[16] R.Aggarwal "Automatic Subspace clustering of high dimensional Data for Data mining Applications".

[16] C.H.Cheng,"Entropy based subspace clustering for Mining Numerical Data" Proc. ACM SIGKDD Int'l Conference Knowledge Discovery and Data Mining (KDD), pp. 84-93, 1999.

[17] P. Kro¨ger, H.-P. Kriegel, and K. Kailing, "Density-Connected Subspace Clustering for High-Dimensional Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 246-257, 2004.

[18] H.Wang,F.Chu "A Fast Algorithm for subspace clustering by pattern similarity".

[19] 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.