An Efficient Approach For Subspace Clustering By Using OCA 3DS Seeker Algorithm

K.Uma Maheswari[1]
Assistant Professor,Department Of Computer Science Engineering
University College Of Engineering(BIT Campus)
Trichy,India
Email: umaravi03@gmail.com

A.Anbarasan[2]
Master Of Engineering, Department Of Computer Science Engineering
University College Of Engineering (BIT Campus)
Trichy,India
Email: anbuvijay26@gmail.com

***Abstract*— In the clustering domain we have identified several problems, which require the mining of actionable subspaces defined by objects and attributes over a progression of time. These subspaces are actionable which means 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. We propose algorithm OCA 3DS Seeker, which uses a hybrid of SVD, optimization algorithm, and 3D frequent item set mining algorithm to mine OCA 3DSs in an efficient and parameter insensitive way. The algorithm is to mine actionable subspace clusters from 3D dimensional data, which are subspaces with high and correlated utilities. The OCA 3DS Seeker provide high efficiency in financial data and is able to discover to significant clusters and find the optimal centroid values to cluster of data with efficient performance and reducible size. We show that our clustering results are not sensitive to the framework parameters. In this paper we show that clusters with higher utilities correspond to higher action ability, and we can able to use our clusters to perform better than one of the most famous value investment strategies.**

Keywords— ***data mining; 3D subspace clustering; singular vector decomposition; financial data mining***

I.INTRODUCTION

Clustering is nothing but analyzing and grouping of datasets from the large databases. It found the datasets based on the data fields. It will recognize the groups of similar objects. Traditionally, a cluster is defined as a subset of objects that is based on the each attribute of the datasets. A more general notion of cluster is a subset of
different applications. These pair-wise dissimilarity measures are often a summary of the dissimilarities across all the attributes.

In the domain area of the users, the objects are grouped together based on their requirements of the users selection basics. Some of the domain areas are astronomy, physics, geology, marketing, etc. Clustering objects are found by the distance between any two objects that are having similar features or value. It can be achieved in a high dimensional data. The high dimensional data means that the similar subset of attributes called subspaces. The entire set of attributes in a high dimensional data is called full space. The datasets can be varied as per the need of the users that are based on their domain area. Moreover, the usefulness of these clusters are used for actions. The actions means that the process of fining high profit that is based on the need of the users.

Traditional clustering algorithms have been successfully applied to low-dimensional data, such as geographical data or spatial data, where the number of attributes is typically small. While objects in some datasets can be naturally described by 3 or fewer attributes, researchers often collect as many attributes as possible to avoid missing anything important. As a result, many datasets contain objects with tens of or even hundreds of attributes. We call such objects high dimensional data. Actions means that the process of fining high profit that is based on the need of the users.

II.RELATED WORKS

A. Mining Customer Value: From Association Rules To Direct Marketing

In this approach[1], the authors K. Wang, S. Zhou, Q. Yang, and J.M.S. Yeung discusses about direct marketing to the customer with the help of Association rule. The direct marketing technique is nothing but making direct communication to the customer sprightly. So the profit can be measured directly. It is used to identify potential customers. Making this direct marketing technique effective, maintain historical database. In this approach estimate directly the profit generated on a customer without estimating the conditional class probability. This method is used for estimating the profit directly. It takes some advantages. First it increases the customer value count and creates a new view for profit estimation. Here use profit estimation with the help of Association rule and pessimistic estimation. Association rule is used to identify the relationship among the variables in large database. This association rule is used for maximize the profit in direct marketing technique. In this approach use two target variables: respond, non-respond.

Association takes more advantages over the local search. In this approach have certain issues. First one is inverse correlation invalidate the probability that is, it gives low rank to valuable customer. Second is, extracting the correlated features from the large data sets. These issues are identified using three approaches. First, extract the features of respond records. Second, build a prediction model for the extracting features to maximize generated profit. Third, remove unwanted features from the model then apply it into the all customer.

In this approach, the association rule increases scalability of correlated features than local search. This approach gave 49% profit using profit estimation using direct marketing with global search of association rule than local search. But in this approach there is no subspace clustering from the datasets. Here we doesn't use conditional probability.

B. Tricluster: An Effective Algorithm For Mining Coherent Clusters In 3d Microarray Data

In this approach[3], the authors L. Zhao and M.J. Zaki explained about a novel, efficient, deterministic, clustering method called Tricluster. They discussed with the gene expression datasets. These kind of datasets are done by initial positioned and overlapping of clusters. They used different parameters values and they focus about the constant or similar values at each dimensions. The main focus was about graph based method for the clustering process. In which each one consisted of time slice. That is the Gene is multiplied by sample matrix. The process is continued to all the similar objects till the work progress in clustering. They used two variables called 'respond' and 'not respond'. In which most of the datasets provide not respond variable only. That is the respond variable will be acknowledged as 5 percent only instead of remaining will provide maximum amount data items. The authors tried to bring 50-50 range of datasets identifications for the both target variables. It means that the respond must reach 50 percent on datasets searching. It addresses certain homogeneity criteria, arbitrary overlapping regions, scaling or shifting expression values, number of samples or time-slices, optionally merge/delete tri clusters. Some features include: 1) For each time slice matrix, find the valid ratio-ranges for all pair of samples, and construct a range multigraph, 2) It Mine the maximal biclusters from the range multigraph, 3) Construct a graph based on the mined biclusters (as vertices) and get the maximal Triclusters, and 4) Optionally, delete or merge clusters if certain overlapping criteria are met.

C .Constrained Locally Weighted Clustering

In this work[5], Hao Cheng. is explained about managing complex, heterogeneous and multidimensional data, makes complex problem in data clustering. To minimize this problem he combined the clusters with independent weighted vector. In this he took constraints with pair level constraints. Then weighted pair were combined into disjoint groups. The combination of the data points one affect another data points. This approach will increase the accuracy than other pervious algorithm techniques. This clustering is accurate, reduce the distance between the data points and the entered point .Here two kinds of approaches were used, one is used for find the distance between data points and next one is used to make division of data points using some constraints. This algorithms conclude the average number of iterations to reach convergence. While Compared with K-Means, the LWC(Locally Weighted Clustering) algorithm converges fairly quickly generate the algorithms to the number of iterations to the clusters. The CLWC algorithm generally took less iteration than K-Means and MPCK-Means. The proper approach is developed to create the weights of dimensions to achieve a better clustering quality of the algorithms.

D. Constraint Based Subspace Clustering

In this work[7] the authors E. Fromont, A. Prado, and C. Robardet discussed about to extend the common framework of bottom-up subspace clustering techniques. In the common high dimensional data, the performance of clustering algorithms are reduced in the traditional manner. Generally the high dimensional data having more noisy data and are irrelevant. To overcome these problems, the current algorithm automatically will find the clusters in the relevant of subset items. In this paper the author deals with the subspace clustering in the bottom-up subspace clustering method. This method integrates with background knowledge and importantly instance level constraints to speed up the subspace clustering. It applied in both density and distance based bottom-up subspace clustering methods. This method will increased not only the efficiency and also focus about the accuracy also. In which the mining involved the text mining and gene expressions analysis. The constraints of the clusters like expected number of clusters, minimum and maximum cluster size and weights for different objects. And the clustering parameters are threshold, distance function. The instance level constraints are must-link and cannot-link. Must-link means that the two objects must be in the same clusters and cannot-link means that the two objects must be in different clusters. here the mining algorithms used called SC-MINER, that is used for subspace clustering under instance level constraints, and it will applied to density based and distance based subspace clustering method. this algorithms achieved more sensitive to possible noise in the data. The constraints directly reduce the sensitivity.

D. Disadvantages

The model can represent the text mining problems easily and directly. However, with the increase of data set size, the vector space becomes high dimensional space, and the computational complexity grows exponentially. Moreover, in many practical applications, completely unsupervised learning is lacking relevant information. On the other hand, supervised learning needs an initial large number of class label information, which requires expensive human labour and time. The limitation of previous algorithm is that it can only tackle the problem when there is only one single type of data objects, i.e. it can only process the homogeneous data set *A.*

III.PROPOSED METHOD

A. Problem Description

OCA 3DS Seeker allows incorporation of user's application domain knowledge on particular area and it is

permits the users to select their preferred objects as centroids, and preferred profitable function to measure the action ability of the clusters. Here we are going to mine the subspace clustering for financial datasets with over a period of years. The new algorithm introduces called OCA 3DS seeker which performs the subspace clustering for the 3D data with profitable of objects on the investment strategies. 3D subspace generation is allowed, as OCA 3DS is in subsets of all three dimensions of the data. Mining OCA 3DS from continuous-valued 3D data is divided into sub problems: 1) pruning of the search space, 2) finding all the subspaces where the objects are homogeneous and have high and correlated utilities, with respect to the centroids, and 3) perform 3D CFI mining to complete OCA 3DS from these subspaces. In proposed system, users should be allowed to incorporate their domain knowledge, by selecting their optimal objects as centroids of the actionable subspace clusters. We denote such clusters as centroid-based, actionable 3D subspace clusters (OCA 3DS), and we also denote utility as a function measuring the profits or benefits of the objects.

IV.SYSTEM ARCHITECTURE

The proposed system is the mining actionable subspace clusters for 3D item sets. Similar to the previous section experimental setup, we embedded actionable subspace clusters in a synthetic dataset D for each experiment. Our visualization framework aids in the analysis of the large sets of subspace clusters mined from a given dataset. It encourages the use of large cluster sets, instead of smaller clusters, and thus retains pattern sensitivity. By clustering subspace clusters it is possible to define summarizations over the whole dataset. It is possible to extract an intrinsic centroid representing the entirety of a cluster of subspace clusters.

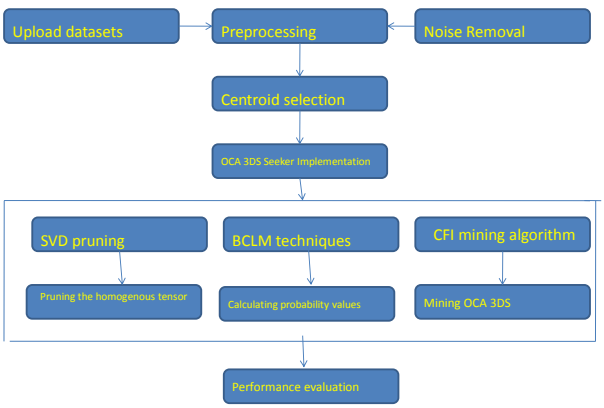

Fig.(a).System Architecture

We can also aggregate subspaces with similar patterns, while retaining interesting or distinct subspaces which would have, in all likelihood, been masked if one is restricted to tuning the clustering parameters. A binary database can be treated as a transactional database, where each transaction represents objects, and each item represents attribute.

A transaction t contains item i if its representative object o has value '1' on attribute a. Recall that the value '1' represents object appearing in the cluster specified by individual attribute, the notion of maximal subspace clusters is equivalent to the closed frequent item set (CFI) of the transactional database. An item set is closed if it has no superset with the same support. The actionable and sequential database is projected into a standard relational database, based on a chosen cluster center c. Note that the projection is per centroid, i.e., we will have one relational database for each cluster centre. In our experiment, we choose the centroids to be objects with utility higher than a parameter minimum. In practice, the domain expert might also select the centroids based on their internal knowledge. The projection is done by setting up an objective function that incorporates the utility, similarity, and correlation of objects

.

A. Upload Datasets

In this module we can upload the datasets. Datasets may be financial datasets or biological datasets. Because we analyse the risks about anyone of these two areas. The datasets consists of financial stock information about any organizations. The data items will contain the financial ratio. The datasets are in the form of three dimensional as object-attribute-time of the products. e.g., the stock-ratio-year data in the finance domain. The file format may be any datasets manageable files like MS-Excel.

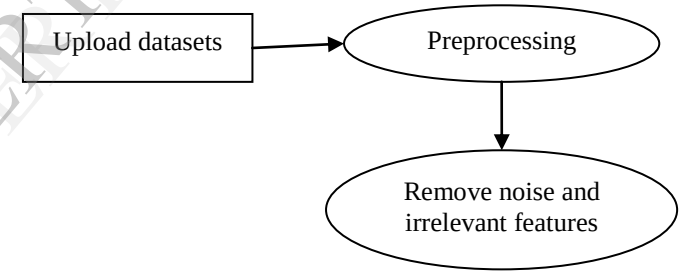

Fig (b) Upload Datasets

B. Pre-Processing

In this module we are going to remove noisy and redundant data. Noisy data means those datas are not in a proper structure or unfilled data. Instance selection is used to find the noisy data which are in the improper manner. Using this selection method we are not able to remove the redundant data. So we move on to the next selection method that is Feature selection. This feature selection method is nothing but to select the relevant feature from the data sets. Irrelevant features are the wasted data. Redundant feature are the repeated information from the predefined data sets.

C. Centroid Selection

In this module we have to find the centroid value from the financial ratio over the financial data sets. Using OCA 3DS Seeker algorithm, we are going to mine the subspace clustering. For mining the subspace clustering we are using centroid value per year. Then find the correct clusters. We analyse the time complexity of each of them, per centroid. Each and every year we choose the optimal centroid to find cluster to analyse maximum risks in financial datasets and also analyse the profit and loss growth.

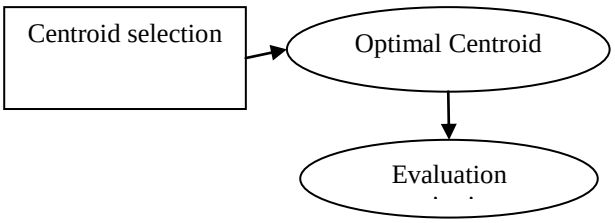

Fig (c) Centroid Selection

D.OCA 3DS Seeker Implementation

a.SVD Pruning

OCA 3DS Seeker uses single value decomposition algorithm to remove the unwanted data items from the search space, which can perform this operation in efficient manner. It remove the content from the uninteresting regions. It performs this work in parameter free way, which means it doesn't consider about it's parameter importance. Then SVD pruning uses the homogeneous matrices to perform their operation and combining the attribute and time dimensions in function expand, instead of other dimensions. It achieves results in less memory usage.

Algorithm: SVDpruning

Shearing the Identical geometric objects Using SVD

Input:
|O| × |A| × |T | Identical geometric objects S
Output:
Sheared Identical geometric objects S
Description:
1: M = expand(S);
2:add mock-up row and column to M;
3:while true do
4: N ← ZeroMeanNormalization(M)
5: U∑V ' ← N
6: u ← PrincipalComp(U);
v ← PrincipalComp(V);
7: calculate thresholds Tu, Tv
8: Shear row i of M if |u(i)| < Tu, 1 <= i <= m
9: Shear column j of M if |v(j)| < Tv, 1 <= j <= n
10: if there is no Shearing then break
11:end while
12:remove mock-up row and column from M;
13:S= shrink(M);

Explanation:

a)Expanding of identical geometric objects:

The given identical geometric objects will be expanded using the expanding function *expand*. That is S(|O|×|A|×|T|) into S(|O|×(|A|×|T|)). For example, a {o1,o2,o3,o4,o5}×(a1,a2}×{t1} into a{o1,o2,o3,o4,o5}×{a1t1,a2t1}. SVD is used to find the covariance matrix of M(Line 1).

The below diagram Fig(a) shows an example of identical elements Matrix M and it's pruned datas after the algorithm has been applied.

We added a Mock-up values to rows and columns that containing all "0" to the Matrices. That is used to check out how far the set of numbers are spread out(Line 2). The sample diagram, after applied the dummy values are shown in fig(b).

b)Purpose of Principal Components:

The Principal Components are used to find the rows and columns of high variance. In which it performs the zero mean normalization that is to get the zero mean normalized matrix N(Line 4). It will be used to calculate the covariance matrices. The zero mean normalization is achieved by taking the average(avj_i) of column values of matrix M. And then

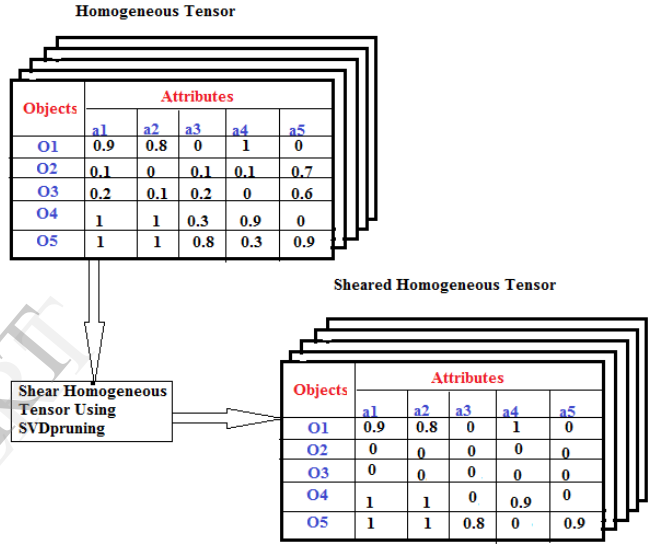

Fig.(d). Calculating and Shearing Homogeneous Geometric Objects

Objects	Features a1,t1	a2,t2	Mock-up
O1	0.83	0.2	0
O2	0.92	0.5	0
O3	1	0.1	0
O4	0.01	0.4	0
O5	0.03	0.2	0
Mock-up	0	0	0

Fig.(e). Data sets after applying Mock-up values

subtracting each entry of M along with their corresponding column mean. The NN' matrix for objects space and N'N is for feature space. The SVD will decompose both covariance matrices(Line 5), as

$NN'=U\sum^2 U'$
$N'N=V\sum^2 V'$

Where,

U -is m ×m orthogonal matrix(Its columns are the eigenvectors of NN').

∑2-is a m×n diagonal matrix with the eigen values on diagonal.

V-is n×n orthogonal matrix(its columns are the eigenvectors of N'N)

m- is a number of rows

n- is a number of columns

We notate u and v as the Principal Component of NN' and2 N'N respectively(Line 6),that is 'u' for object spaces and 'v' for feature spaces. Calculate the threshold for the principal components for u and v. i.e. Tu and Tv(Line 7).Then shear the rows and columns of M with respect to i and j.That is the rows and columns within the principal component must be less than the threshold of the same principal components. i.e. Tu and Tv(Line 8,9). If there is no shearing on datas then finish the process(Line 10,11).Then remove all the Mock-up values from matrix M(Line 12).Finally shrink the matrix M and assign it to S.

b. BCLM Implementation (Bound-Constrained Lagrangian Method)

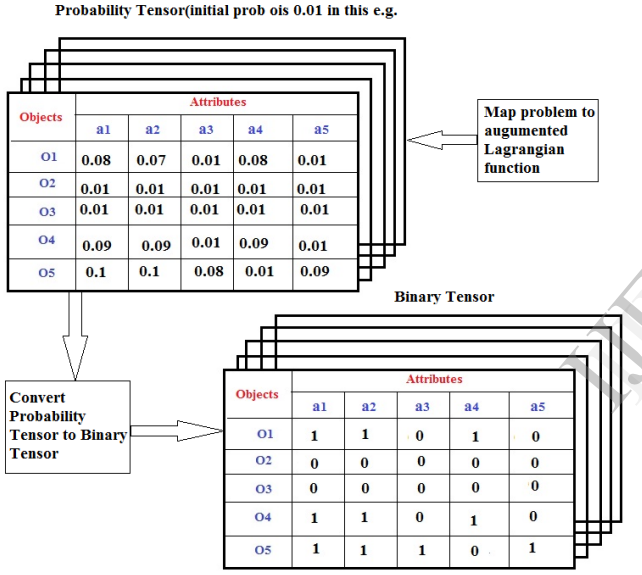

Fig.(f). Calculating the probabilities of the values

OCA 3DS Seeker 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. Here the parameter is insensitive. The augmented Lagrangian Multiplier Method is used to select the maximum financial ratio value from the datasets. This approach is a robust on clusters because the objects are in the form of movable format. It converts the financial ratio value those are having more than the threshold range based on centroid into binary 1's else binay 0's.

c. 3D CFI Mining

OCA 3DS Seeker uses the 3D closed frequent itemset mining method to efficiently mine subspace clusters, based on the score of the objects in the subspaces. We then use efficient 3D closed pattern mining algorithms to efficiently mine sub cuboids from the BCLM resultant values into binary 1's only. These values are clustered items with better profit and efficiency which correspond to the OCA 3DS.

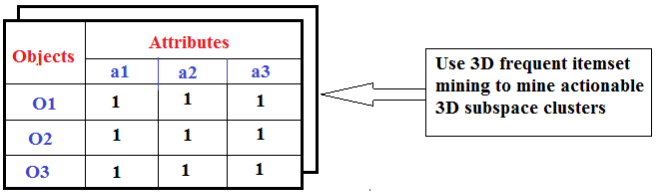

Fig.(g).Example of an Actionable 3D Subspace Clustering

E. Performance Evaluation

We can show that OCA 3DS Seeker will have 90 percent higher profit/risk ratio than the existing approaches in financial data. The evaluation criteria includes efficiency analysis and parameter sensitivity analysis, quality analysis of the clusters are mined by OCA 3DS seeker on the application of stock market.

F.Advantages

The actionable subspace clusters have the following advantages.

- In this the financial ratios are having more utility, so that the action suggested by the cluster is profitable or beneficial to the user.
- Analysis the report accurately
- Time consuming process
- Accuracy can be upto 90%
- Handle high dimensional datasets with effectiveness.
- The utilities of the objects are related to each other, so that these objects with homogeneous values and high utility do not occur together by chance.

V.DATA SETS USED

We used synthetic data sets in our experiments. In this kind of data set D, the values of objects are randomly valued from 0 to 1.And then the utilities will be valued from -1 to 1.We randomly used 10 OCA 3DS Seeker, each having 15-25 objects, 5 to 10 attributes and 5 to 10 time stamps in the data sets D. To confirm with the homogeneity with the objects in each cluster, we set a maximum difference allowance named as diff on its objects. In second attempt of synthetic data sets D used, that contains 1,000 objects ,10 attributes and 10 time stamps. The sample utility of the object in the clusters at least 0.5 and the diff is at most 0.1.So totally the sample data volume will be 1 million.

VI. EFFICIENCY ANALYSIS OF OCA 3DS SEEKER

We investigated the computation costs of OCA 3DS seeker on the continuous valued 3D Data set D.We mainly focused on the similarly on the attribute dimension not with the experiment on the time dimension. We had the efficiency comparison of OCA 3DS Seeker with the Mining Actionable Subspace Clustering(MASC)[9] only. Because it is the only algorithm for centroid-based 3D subspace clustering. And also we made the comparison along with the We also compared OCA 3DS Seeker with STATPC and MIC, because these two algorithms are parameter insensitive and the default parameter

setting can be used. We did not compare with the other algorithms, as their efficiency depends on their parameter settings. Given that both OCA 3DS Seeker and MASC are run per centroid, Totally we ran 10 centroid vales for each data set D. And then make the average of their running time to obtain the results of the efficiency.

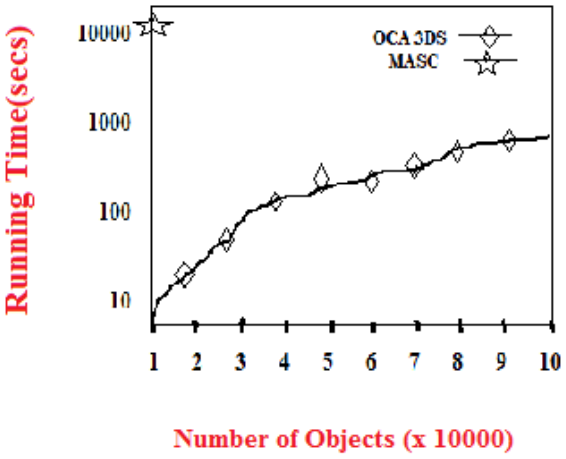

Fig.(h). Running time of the OCA 3DS Seeker with MASC

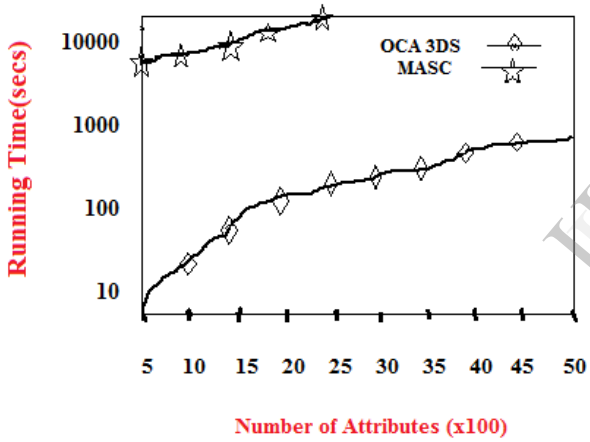

Fig.(i). Running time of OCA 3DS Seeker with polynomial function of the number of attributes

First, we varied the number of objects from 10,000 to 100,000 and fixed the number of time stamps. i.e. the year value, and number of attributes at 10.The result of medium size 3D data set from 1 to 10 million values. The below figure(h) shows the running time of the OCA 3DS Seeker with the STATPC[11] and MIC[10]. In which Both the STATPC and MIC could not finish running after 6 hours. next, we changed the number of attributes from 500 to 4,000 and fixed the number of objects at 1,000 and number of time stamps at 10. So the resulting of the of the medium sized 3D dataset values were from 5to 40 millions. Fig. (i) presents the results, which also shows that the running time of OCA 3DS Seeker is a polynomial function of the number of attributes. Again, STATPC and MIC could not finish running after 6 hours.

Both experiments show that OCA 3DS Seeker took at most 500 seconds to finish the mining task, which is reasonable for real world applications containing medium-sized data sets. For MASC, we can see that it is orders of magnitude slower than OCA 3DS Seeker. This is happening due to two reasons; At first, for each centroid, the MASC needs to run its optimization algorithm |A| times. But OCA 3DS Seeker only needs to run its optimization algorithm once. Second, MASC does not prune the data set, while OCA 3DS Seeker does. STATPC and MIC are slower than OCA 3DS Seeker as they mine the complete set of subspace clusters, while OCA 3DS Seeker is run per centroid.

VII.CONCLUSION

We have performed the subspace clustering by using OCA 3DS Seeker algorithms with a efficient and optimal manner because of selection of optimal centroid values on the 3D financial data set values. In future we can implement 3D subspace clustering by using both the optimal and fixed centroid values.

VIII.REFERENCES

[1] 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.

[2] 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.

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

[4] 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,.

[5] H. Cheng, K.A. Hua, and K. Vu, "Constrained Locally Weighted Clustering," Proc. VLDB Endowment, vol. 1, no. 1, pp. 90-101, 2008
[6] 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..

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

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

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

[11] G.Moise and J. Sander, "Finding Non-Redundant, StatisticallySignificant Regions in High Dimensional Data: A Novel Approachto Projected and Subspace Clustering," Proc. 14th ACM SIGKDDInt'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 2008.