**International Journal On Engineering Technology and Sciences-IJETS™**
**ISSN(P):** 2349-3968, **ISSN(O):** 2349-3976
**Volume I, Issue I, May-2014**

# Determining and Exploring 3Dimensional Data using Value Decomposition

***P.Thiyagarajan[1], R.Aarthi[2]***
*PGScholar[1] , Assistant Professor[2], Nandha College of Technology, Erode*
thiyagu33326@gmail.com[1], aarthikalai@gmail.com[2]

**Abstract**: *Clustering a large sparse and large scale data is a open research in the data mining. To discover the significant information through clustering algorithm stands inadequate as most of the data finds to be non actionable .Existing clustering technique is not feasible to time varying data in high dimensional space. Hence Subspace clustering will be answerable to problems in the clustering through incorporation of domain knowledge and parameter sensitive prediction. Sensitiveness of the data is also predicted through thresholding mechanism. The problems of usability and usefulness in 3D subspace clustering are very important issue in subspace clustering. Also determining the correct dimension is inconsistent and challenging issue in subspace clustering .In this thesis, we propose Centroid based Subspace Forecasting Framework by constraints is proposed, i.e. must link and must not link with domain knowledge. Unsupervised Subspace clustering algorithm with inbuilt process like inconsistent constraints correlating to dimensions has been resolved through singular value decomposition. Principle component analysis is been used in which condition has been explored to estimate the strength of actionable to be particular attributes and utilizing the domain knowledge to refinement and validating the optimal centroids dynamically. An experimental result proves that proposed framework outperforms other competition subspace clustering technique in terms of efficiency, Fmeasure, parameter insensitiveness and accuracy.*

## I.INTRODUCTION

In this work, proposed and analysed the Subspace clustering through forecasting framework for value decomposition in high dimension actionable data. aims to find groups of similar objects and due to its usefulness, it is popular in a large variety ofdomains, such as geology, marketing and etc.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. The objects in a group need only to be similar on a subset of attributes .The high-dimensional data sets in these domains also potentially change over time. We define such data sets are defined by three-dimensional (3D) data sets which can be generally expressed in the form of object-attribute-time. Such data sets are used to finding subspace clusters per time

This stamp may produce a lot of spurious and arbitrary clusters. It is also 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. Such patterns are called actionable patterns. 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 theirpreferred objects as centroids. The cluster objects are similar to the centroids. In this paper, we identify real-world problems, which actionability and users domain knowledge via centroids. Existing 3D subspace clustering algorithms are inadequate in mining actionable 3D subspace clusters. We propose mining Centroid-based, actionable 3D Subspace clusters with respect to a set of centroids, to solve the above issues.

CBSF Clustering 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 CBSF Clustering is in subsets of all three dimensions of the data. Mining CBSFSs from continuous-valued 3D data is nontrivial. It is necessary to breakdown

this complex problem into sub problems: 1) pruning of the search space, 2) finding subspaces in the objects where homogeneous are correlated utilities 3) mining CBSFSs from these subspaces. We propose a novel algorithm.

CBSF clustering uses SVD to prune the search space.That are efficiently prune the uninteresting regions.This approach is parameter free.

- CBSF Clustering uses augmented Lagrangian multiplier method to score the objects in subspaces where they are homogeneous and have high and correlated utilities.
- CBSF clustering uses the state of the art 3D frequent itemset

## II.PROPOSED CONTRIBUTIONS

We identify the need to mine actionable data through subspace clustering , which are clusters of objects that suggest profits or benefits to users. The users are allowed to incorporate their domain knowledge by selecting their preferred objects as centroids of the clusters. We proposed algorithm Forecasting algorithm which used a hybrid of SVD optimization algorithm and 3D frequent itemset mining algorithm. We conduct a comprehensive list of experiments to verify the effectiveness of value decomposition technique and to demonstrate its strengths over existing approaches:

- Robustness. Correct clusters are found using CBSF (centroid based subspace forecast) clustering, even with 20 percent perturbation in the data.
- Parameter insensitivity. Correct clusters are found across diverse settings of CBSF tuning parameters.
- Effectiveness. CBSF clustering has on average 180 percent higher accuracy in recovering embedded clusters than current subspace clustering algorithms.
- Efficiency. CBSF is at least 2 orders of magnitude faster than the other centroid-based subspace clustering algorithm.
- AppliCBSFions on real-world data. We show that CBSF clustering 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.

## III.METHODS

1. Dataset Pre-processing
2. Designing the unsupervised Constraints based on domain knowledge.
3. Establishing Centroid based 3D subspace clustering
4. A .Singular Value Decomposition of actionable centroids
   a. Numerical optimization of cluster values
   b. Frequent Item set Mining
5. Establishing the forecasting technique to optimal centroids
6. Selection of Dimensions through Actionable Weight using Principle component Analysis.
7. Performance Comparison

### 1. Dataset Processing

In this module we going to build a synthesis dataset for performing for the processes mentioned in the following modules. This module is contains high dimensional data as a synthesis dataset as its contains more information with several attributes along huge records in difference time factors to analyse for providing accurate predictions in future cases.

### 2. Designing the unsupervised Constraints based on domain knowledge.

Many machine learning tasks require similarity functions that estimate likeness between observations. Similarity computations are particularly important for clustering and record linkage algorithms that depend on accurate estimates of the distance between data points. However, standard measures such as Euclidean distance is the most common use of distance, examines the root of square differences between coordinates of a pair of objects.

### 3. Establishing Centroid based 3D subspace clustering

#### i) Singular Value Decomposition of actionable centroids

In this Centroid-based, Actionable 3D Subspace clustering (CATs), the high-dimensional and continuous- valued tensor is a difficult and time-consuming process. Hence, it is vital to first remove regions that do not contain CATs.

A simple solution is by removing values that are less than a threshold, but it is impossible to know the right threshold. Hence, we propose mechanism to efficiently prune tensor in a parameter-free way, by using the variance of the data to identify regions of high homogeneity values.

**ii) Numerical optimization of cluster values**

Here we use the homogeneous tensor with the utilities of the objects to calculate the probability of each value of the data to be clustered with the Centroid. After calculating the probabilities of the values, we binarize the values that have high probabilities.

**iii) Frequent Item set Mining**

One of the most common approaches to mining frequent patterns is the apriori method and when a transactional database represented as a set of sequences of transactions performed by one entity is used, the manipulation of temporal sequences requires that some adaptations be made to the apriori algorithm.

The most important modification is on the notion of support: support is now the fraction of entities, which had consumed the itemsets in any of their possible transaction, i.e. an entity could only contribute one time to increment the support of each itemset, beside it could had consumed that itemset several times.

After identifying the large itemsets, the itemsets with support greater than the minimum support allowed, they are translated to an integer, and each sequence is transformed in a new sequence, whose elements are the large itemsets of the previous-one.

The next step is to find the large sequences. For achieve this, the algorithm acts iteratively as apriori: first it generates the candidate sequences and then it chooses the large sequences from the candidate ones, until there are no candidates. One of the most costly operations in apriori-
based approaches is the candidate generation.

**4. Establishing the forecasting technique to optimal centroids**

The need of analyzing and grouping of data is required for better understanding and examination. This can be solved by using the clustering technique which groups the similar kind data into a particular cluster. One of the most commonly and widely used clustering is K-Means clustering because of its simplicity and performance.

**5. Selection of Dimensions through Actionable Weight using Principle component Analysis.**

If the dataset used is large, then the performance of K-Means will be reduced and also the time complexity is increased. To overcome this problem, this method focuses on altering the initial cluster Centroid effectively, for this purpose, Principal Component Analysis (PCA) is used here.

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. K-means clustering is a commonly used data clustering for unsupervised learning tasks. Here we prove that principal components are the continuous solutions to the discrete cluster membership indicators for K-means clustering. The experimental result shows that the proposed technique results in better accuracy and also the time complexity is reduced.

6. **Performance comparison**

In this experiment, an analysis is done on the behaviour of the clusters of K-Means and PCA algorithms. Experiments on dataset PCA provides an effective Clustering solution for the K-means clustering.

## IV.CONCLUSION

Mining actionable 3D subspace clusters from continuous valued 3D (object-attribute-time) data is useful in domains ranging from

finance to biology.But this problem is nontrival as it requires input of users‟ domain knowledge,clusters in 3D subspaces,and parameter insensitive and efficient algorithm.

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 synthetical 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 percen better than the next best competitor in the return/riskratio.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] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When Is Nearest Neighbor Meaningful?” Proc. Seventh Intl Conf. Database Theory (ICDT), pp. 217-235, 1999.

[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, 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 Intl Conf. Extending Database Technology:Advances in Database Technology (EDBT), pp. 70-87, 2002.

[5] K. Wang, S. Zhou, Q. Yang, and J.M. Yeung, “Mining CustomerValue: From Association Rules to Direct Marketing,” Data Mining Knowledge Discovery, vol. 11, no. 1, pp. 57-79, 2005.

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

[7] B. Graham, The Intelligent Investor: A Book of Practical Counsel.Harper Collins Publishers, 1986.

[8] 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.

[9] J.F. Swain and L.M. Gierasch, “The Changing Landscape of Protein Allostery,” Current Opinion in Structural Biology, vol. 16, no. 1, pp. 102-108, 2006.

[10] G. Buhrman et al., “Allosteric Modulation of Ras Positions Q61 for a Direct Role in Catalysis,” Proc. Nat‟l Academy of Sciences USA,vol. 107, no. 11, pp. 4931-4936, 2010.

[11] P.Bradley & U.Fayyad, “Refining Initial Points forK-Means Clustering”. In Proceedings of the 15th ICML, 91-99, Madison, WI, (1998).

[12] G.P.Babu & M.N.Marty, “Clustering with evolutionstrategies Pattern Recognition”, 27, 2, 321-329, (1994).

[13]R.Duda&P.Hart, ”Pattern Classification and SceneAnalysis”. John Wiley & Sons, New York, NY, (1973).

[14] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda,G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach,D.J. Hand, and D. Steinberg, “Top 10 Algorithms in Data Mining,”Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, (2008).

[15] Q. Yang and X. Wu, “10 Challenging Problems in Data MiningResearch,” Int‟l J. Information Technology and Decision Making, vol. 5,no. 4, pp. 597-604, (2006). [16] G.P.C. Fung, J.X. Yu, H. Lu, and P.S. Yu, “Text Classificationwithout Negative Examples Revisit,” IEEE Trans. Knowledge and DataEngineering, vol. 18, no. 1, pp. 6-20, (Jan 2006).

[17] H. Al Mubaid and S.A. Umair, “A New Text CategorizationTechnique Using Distributional Clustering and Learning Logic,”IEEE Trans. Knowledge and Eng., vol. 18, no. 9, pp. 1156-1165, (Sept 2006).

[18]Gregory. Piatetsky, “Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from university to business and analytics. Data Mining Knowledge

Discovery (2007)15:99– 105,(2007).
[19] Hans-Peter.Kriegel, Karsten M .Borgwardt &Peer. Kröger, "Future trends in data mining". Data Mining Knowledge Discovery 15:87–97, (2007).
[20] Qi.Luo, "Advancing Knowledge Discoveryand Data Mining" Knowledge Discovery and Data Mining, WKDD (2008).
[21]Gary M.Weiss, Bianca.Zadrozny, Maytal.Saar-Tsechansky, "Guest editorial: special issue onutility-based data mining". Data Mining Knowledge Discovery 17:129–135, (2008).
[22]Ben G.Weber &Michael.Mateas , "A Data Mining Approach to Strategy Prediction " 978-1- 4244-4815, IEEE, (2009).
[23]Sufal.Das &Banani.Saha,"Data Quality Mining using Genetic Algorithm". International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (2), (2009).
[24]Atul. Kamble,"Incremental Clustering in Data Mining using Genetic Algorithm". International Journal of Computer Theory and Engineering, Vol.2, No. 3, (June 2010). [25] Murat.Kantarcıoglu, Bowei.Xi &Chris.Clifton, "Classifier evaluation and attribute selection against active adversaries" Data Mining Knowledge Discovery 22:291–335, (2011).
[26] R.Bouckaert, "Naive Bayes Classifiers That Perform Well with Continuous Variables, Lecture Notes in Computer Science". Volume 3339, Pages 1089 – 1094,(2004).
[27] L. A.Breslow,&D.W.Aha, "Simplifying decision trees: A survey". Knowledge Engineering Review 12: 1–40, (1997).