ISSN:0975-9646

An Efficient Density Conscious Subspace Clustering Method using Top-down and Bottom-up Strategies

[1]M. Suguna [2] Dr. S. Palaniammal
[1]*Assistant Professor, Department of Computer Science,*
Dr. GRD College of Science, Coimbatore.
[2]*Professor & Head, Department of Science & Humanities,*
Sri Krishna College of Technology, Coimbatore.

Abstract **- Clustering high dimensional data is an emerging research field. Most clustering technique use distance measures to build clusters. In high dimensional spaces, traditional clustering algorithms suffers from a problem called *"curse of dimensionality"*. Subspace clustering groups similar objects embedded in subspace of full space. Recent approaches attempt to find clusters embedded in subspace of high dimensional data. Most of the previous subspace clustering works discovers subspace clusters, by regarding the clusters as regions of higher densities. The regions are identified dense if its density exceeds the density threshold. As the cluster densities varies in different subspace cardinalities, it suffers from a problem called *"density divergence problem"*. We follow the basic assumptions of previous work DENCOS. It is found that varying region densities are used to overcome density divergence problem. All previous approaches are based on bottom-up method. In this paper a novel data structure is used which works on both bottom-up & top-down fashion. Performance results of this new novel data structure shows very good results and the efficiency outperforms the previous works.**

Key Words **- Subspace Clustering, High dimensional data, Mining frequent patterns, Top down, Bottom Up**.

I. INTRODUCTION

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [10], [11]. Clustering is very useful technique for the discovery of data distraction and patterns in underlying data. The goal of clustering is to discover the dense and sparse regions in the dataset. Cluster detection is based on similarity between objects, typically measured with respect to distance functions. In high dimensional space, effects attributed to the "*curse of dimensionality*" [12], [16] are known to break traditional clustering algorithms. Meaningful clusters cannot be detected as distances are increasingly similar to growing dimensionality.

A common approach to overcome this difficulty is by reducing data dimensionality by using techniques such as feature transformation and feature selection. Feature transformation methods [10], [9], such as PCA & SVD transform the data onto a smaller space while generally preserving the original relative distance between objects. They summarize data by creating linear combinations of the attributes and may discover hidden structures in the data, even after transformation. Thus the transformed features are difficult to interpret, making the clustering results less useful.

Another way of tackling curse of dimensionality is to try to remove some of the dimensions. Feature selection [14], [10] is commonly used for data reduction by removing irrelevant or redundant dimensions. Only particular subspaces are selected to discover clusters. However in many real datasets clusters may be embedded in varying subspaces, thus in feature selection the information about data points resided in varying subspaces is lost.

Subspace clustering [4], [13], [15] is the extension to feature selection technique. 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 dataset. Most of the previous subspace clustering works discovers the subspace clusters by regarding clusters as regions of higher densities than their surroundings in a subspace. High density regions are identified by introducing a *density threshold*. The regions are identified dense if its density exceeds the density threshold. As the cluster densities vary in different subspace cardinalities, a particular density threshold cannot be used to find clusters in all subspace, since the identification of high density regions lacks of considering a critical problem called *density divergence problem*.

The density divergence problem [6] refers to the phenomenon that the cluster densities vary in different subspace cardinalities. Note that as the no. of dimensions increase, data points are spread out in a large dimensional space such that they will be more sparsely populated in nature. This phenomenon implies that finding clusters in higher subspaces should be with a lower density requirement (otherwise we may lose true clusters in such situations), thus showing the existence of density divergence problem. A reasonable consideration is to take densities in different subspace cardinalities. To achieve this, innovative algorithm is devised [6] to adaptively determine the density thresholds for different subspace cardinalities.

We can follow the procedure proposed in [6] to overcome density divergence problem. Motivated by this idea, we propose a novel data structure that works based on both top-down and bottom-up strategy. A novel FP-tree [7]

is constructed to store crucial information of the dataset. The basic idea is by transforming the problem of identifying *dense units* in subspace clustering into a similar problem of discovering *frequent itemsets* in association rule mining. Thus we construct the compact structure by storing the complete information of the dense units satisfying different thresholds in different subspace cardinalities such that dense units can be discovered from this structure efficiently.

LAYOUT OF THE PAPER

The rest of the paper is organized as follows, Section III, some related work in subspace clustering. In section IV, we present the proposed algorithm, in section V experimental results are reported and finally section VI concludes the paper.

II. LITERATURE REVIEW

Subspace clustering [15] is an extension of traditional clustering algorithms; it aims to find clusters embedded in subspace of high dimensional dataset. Subspace clustering algorithms can be divided into 2 categories: Grid and Density based approaches. CLIQUE [1] algorithm was one of the first subspace clustering algorithms. The algorithm combines density and grid based clustering and uses an apriori style technique to find clusterable subspaces. Clusters are defined as regions of higher densities; the regions are identified as dense if it exceeds a density threshold. Subspace clusters are formed by combining connected dense units. However by utilizing single density threshold for all subspace cardinalities; it is difficult for CLIQUE to discover high-quality clusters in all subspaces. This is due unit densities varies in different subspace cardinalities.

ENCLUS [2] is a subspace clustering method based heavily on the CLIQUE algorithm. ENCLUS introduces the concept of '*subspace Entropy*'. The algorithm is based on the observation that a subspace with clusters typically has lower entropy than a subspace without clusters. ENCLUS uses the same APRIORI style, bottom-up approach as CLIQUE to mine significant subspaces such that the density divergence problem is faced. MAFIA [3] is another extension of CLIQUE that uses an adaptive grid based on the distribution of data to improve efficiency and cluster quality, also introduces parallelism to improve scalability. However the main drawback is using a global density threshold for all subspaces so density divergence problem is faced.

In DENCOS [6] different density thresholds is utilized to discover the clusters in different subspace cardinalities to cop up with density divergence problem. Here the dense unit discovery is performed by utilizing a novel data structure DFP-tree (Density FP-tree) which is constructed on the data set to store the complete information of the dense units. To tackle the density divergence problem, this algorithm devise a novel subspace clustering model to discover the clusters based on the relative region densities in the subspaces. DENCOS suffers from recursive construction and materialization of conditional frequent pattern trees which dramatically increases the mining cost.

SUBCLU [5] overcomes the limitations of grid-based approaches like the dependence on the positioning of the grids. Instead of using grids the DBSCAN [17] cluster model of density-connected sets is used. SUBCLU is based on a bottom-up, greedy algorithm to detect the density-connected clusters in all subspaces of high-dimensional data. SUBCLU will also suffer from density divergence problem. Thus it will have difficulties in identifying the core objects in different subspace cardinalities by using the same parameter setting, thus resulting in poor clustering results.

III. PROPOSED METHODOLOGY

A. Preprocessing Dataset

The dataset is preprocessed by transforming each d-dimensional data point into d 1-dimensional data point. Each d 1-dimesional unit is treated as item. The database is scanned to compute the set of all frequent items with density threshold and order them in the density threshold descending order. Then creating the root of the tree, another database scan is performed to build FP-tree.

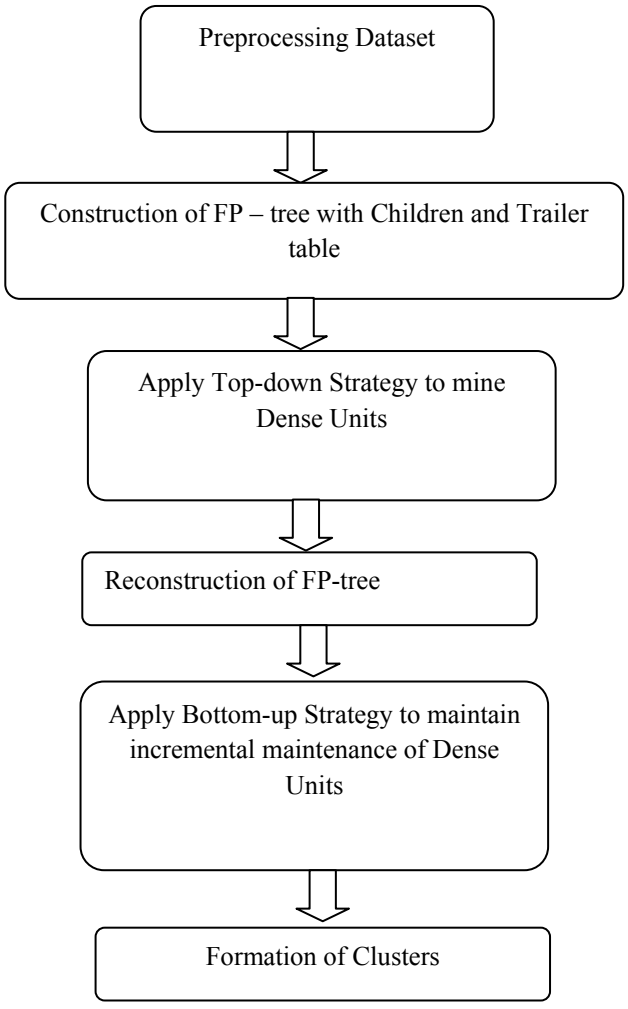

Fig 1. System Flow Diagram

B. Construction of FP-tree with Children and Trailer Table

After scanning the first transaction of the DB, the first branch of the tree with the frequent items is constructed. Those items that are deleted will be inserted into new database named DB', at the same time the transaction is added to the trailer table and let the new transaction's trail of node link pointing to the last node in that path. If all the items in a transaction are frequent, then the items are added to the FP-tree and that transaction can be deleted directly by not adding any item to the new database DB'. In a transaction if all the items are frequent then no need to add that transaction to the trailer table.

C. Top-down Strategy for Mining Dense Units

In this section a novel algorithm is presented to mine dense units FP-tree. A divide and-rule strategy of Top-down method is adopted, where Prefix is used to identify a subtree. Assuming i_j or i_k for a single item, a or b for an itemset which contains multi items or empty. The Top-Down algorithm adopts an Apriori heuristic (however it doesn't bring the candidate generation.). An itemset can ai_ji_k never be dense if the item set represented by ai_j is not dense. When the itemset ai_j is not dense, no need to test the whole ai_ji_k-prefixes ($j< k \leq n$).

D. Reconstruction of FP-tree

FP-tree is reconstructed when the density threshold is changed. Let τ is the old density threshold, and F is the set of all dense units that are mined. When the density threshold τ is changed to τ', there are two probabilities: if $\tau' > \tau$, some frequent units are not frequent any more, otherwise, some infrequent units will be frequent. Setting the new frequent unit set as F'. Because τ' is smaller than τ, some units that were not frequent will be frequent. But the information about new frequent unit is not in the original FP-Tree. After the reconstruction, the new FP-tree will contain all information about new frequent units.

E. Bottom-up Strategy to maintain incremental maintenance of dense units

Based on the FP-tree constructed by step-2 or reconstructed by step-4, the algorithm for mining this tree can be performed. For first time mining $\tau'=0$, then call top-down algorithm directly. Otherwise τ is compared with τ'. If $\tau< \tau'$, the new frequent dense units F' can be constructed. If $\tau> \tau'$, some units that were not frequent may become frequent. By the FP-tree Reconstruction algorithm, the new FP-tree that contains new frequent units is achieved. Only all frequent patterns containing new frequent items need be computed. The process of mining is not from scratch. All work done can be recycled substantially.

F. Cluster Formation

After the dense units are mined, the clusters are being formed by merging the dense units that have common faces. Clusters are being formed in all different subspaces with different density thresholds.

IV. Experimental Results

In this section, we evaluate the performance of our method. The proposed algorithm is compared with the existing algorithm DENCOS [6] and CLIQUE [1] in order to show the advantages of our algorithm. From our proposed algorithm we know that it can avoid generating conditional pattern for each item, so the time and cost of mining is reduced. We have tested the algorithm with adult data set and thyroid disease dataset.

Real Data sets

Dataset	No. of Data Points	Attributes
Adult	32561	14
Thyroid	18152	21

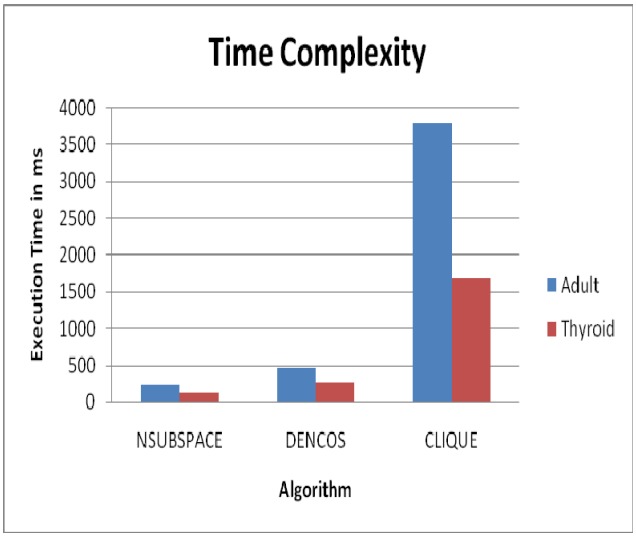

Fig.1: Runtime efficiency on Real data sets.

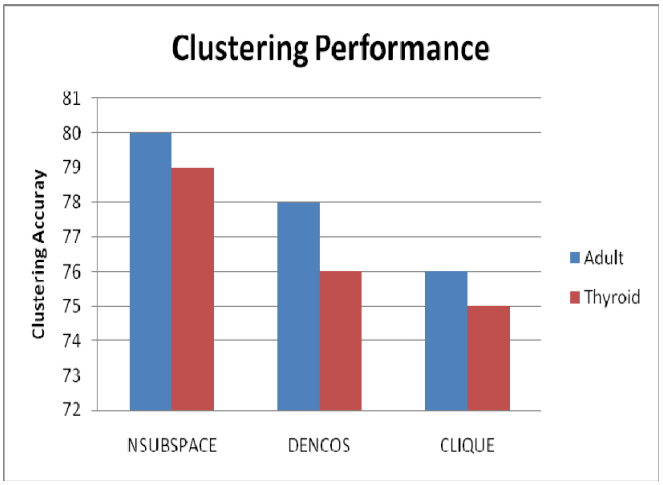

Fig.2: Clustering efficiency of Proposed algorithm on several other algorithms.

The performance of the proposed algorithm is presented in fig.1, fig.2. The graph shows that the proposed algorithm outperforms with DENCOS and CLIQUE in both time and clustering accuracy.

V. CONCLUSION

We have proposed a novel data structure DFFP for storing curical information about frequent patterns. The proposed algorithm works in both top-down and bottom-up fashion. All previous subspace clustering methods works on bottom-up fashion. In previous work [6], FP-growth like approach (DFP-tree) is used which does not bring candidate generation. However, the cost of recursively constructing each frequent item's conditional pattern tree is high. In order to overcome this problem depth first method is adopted. The proposed method is an efficient, incremental updating frequent pattern mining technique for maintenance of the frequent patterns. This method uses FP-tree with children table and trailer table to avoid repeated scanning, reconstructing, and computing. As shown by experimental results, proposed methodology is efficient and scalable for mining frequent patterns and in order of magnitude faster than DFP-tree.

REFERENCES

[1] R.Agrawal, J.Gehrke, D.Gunopulos, andP.Raghavan, *"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications,"* Proc. ACM SIGMOD Int'l Conf. Management of Data, 1998.

[2] C.H. Cheng, A.W. Fu, and Y. Zhang, *"Entropy-Based Subspace Clustering for Mining Numerical Data,"* Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 1999.

[3] H.S. Nagesh, S. Goil, and A. Choudhary, *"Adaptive Grids for Clustering Massive Data Sets,"* Proc. First SIAM Int'l Conf. Data Mining (SDM), 2001.

[4] K. Kailing, H.-P. Kriegel, and P. Kroger, *"Density-Connected Subspace Clustering for High-Dimensional Data,"* Proc. Fourth SIAM Int'l Conf. Data Mining (SDM), 2004.

[5] Yi-Hong Chu, Jen-Wei Huang, †Kun-Ta Chuang, and Ming-Syan Chen *"On Subspace Clustering with Density Consciousness"*.

[6] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, Member, IEEE, and Ming-Syan Chen, Fellow, IEEE, *" Density Conscious Subspace Clustering of High Dimensional Data"*, IEEE Transactions on knowledge and data Engineering, 2010.

[7] Qunxiong Zhu and Xiaoyong Lin, *"Top-Down and Bottom-Up Strategies for Incremental Maintenance of Frequent Patterns"*, School of Infromation Science and Technology, 2007.

[8] Han, J., Pei, J., Yin, Y.: *"Mining Frequent Patterns without Candidate Generation"*. In: Intl. Proc. of the 2000 ACM SIGMOD, Dallas, pp. 1–12 (2000)

[9] Pang – Ning Tan, Michael Steinbach, Vipin Kumar, *"Introduction to Data Mining"*, Pearson Education, 2006.

[10]. Jiawei Han and Michaline kamber, *"Data Mining Concepts and Techniques"*, Morgan Kaufmann Publishers, 2nd Edition.

[11] Margaret H.Dunham, *"Data Mining Introductory and Advanced Topics"*, Pearson Education, 2006.

[12] C.C. Aggarwal, A. Hinneburg, and D. Keim, *"On the Surprising Behavior of Distance Metrics in High Dimensional Space,"* Proc. Eighth Int'l Conf. Database Theory (ICDT), 2001.

[13] Ren´ e Vidal, Johns Hopkins University *"A Tutorial on Subspace Clustering"*.

[14] A. Blum and P. Langley, *"Selection of Relevant Features and Examples in Machine Learning,"* Artificial Intelligence, vol. 97, pp. 245-271, 1997.

[15] L. Parsons, E. Haque, and H. Liu, *"Subspace Clustering for High Dimensional Data: A Review,"* ACM SIGKDD Explorations Newsletter, vol. 6, pp. 90-105, 2004.

[16] A. Hinneburg, C.C. Aggarwal, and D. Keim, *"What is the Nearest Neighbor in High Dimensional Spaces?"* Proc. 26th Int'l Conf. Very Large Data Bases (VLDB), 2000.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,"* Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD), 1996.