On Subspace Clustering with Density Consciousness

Yi-Hong Chu, Jen-Wei Huang, †Kun-Ta Chuang, and Ming-Syan Chen
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, ROC
†Graduate Institute of Communication Engineering
National Taiwan University
Taipei, Taiwan, ROC
E-mail: {yihong, jwhuang, †doug}@arbor.ee.ntu.edu.tw, mschen@cc.ee.ntu.edu.tw

ABSTRACT

In this paper, a problem, called "the density divergence problem" is explored. This problem is related to the phenomenon that the densities of the clusters vary in different subspace cardinalities. We take the densities into consideration in subspace clustering and explore an algorithm to adaptively determine different density thresholds to discover clusters in different subspace cardinalities.

Categories and Subject Descriptors: H.2.8 [Database Applications]: Data mining

General Terms: Algorithms, Management

Keywords: Subspace clustering, Density divergence problem

1. INTRODUCTION

Among recent studies on high-dimensional data clustering, subspace clustering is the task of automatically detecting clusters in subspaces of the original feature space. Most of previous works [1][2][4][5] adopt the density-based approach, where clusters are regarded as regions of high density in a subspace that are separated by regions of lower density.

However, a critical problem, called "*the density divergence problem*" is ignored in mining subspace clusters such that it is infeasible for previous subspace clustering algorithms to simultaneously achieve high precision and recall[1] for clusters in different subspace cardinalities. *"The density divergence problem"* refers to the phenomenon that the cluster densities vary in different subspace cardinalities. Note that as the number of dimensions increases, data points are spread out in a larger 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 the density divergence problem. Due to the requirement of varying density thresholds for clusters in different dimensions, it is challenging in subspace clustering to simultaneously achieve high precision and recall for clusters in different subspace cardinalities. More explicitly, since previous subspace clustering algorithms [1][2][4][5] identify the high-density regions (clusters) in all subspaces with a single density threshold, the trade-off between recall and precision will be inevitably faced. The density threshold should be set low enough for the higher dimensional clusters being discovered with high recall. However, for the clusters in lower subspace, the low threshold may lead the low-density regions around the clusters to be discovered as dense ones, thus resulting in decreased precision of the lower dimensional clusters.

Clearly, the trade-off between precision and recall in previous subspace clustering, which is incurred by the "density divergence problem," solely depends on the determination of the density threshold. However, it is quite subtle to set an appropriate density threshold, and the parameter determination is fully left unsolved to users, thus degrading the applicability of subspace clustering. A reasonable consideration is to take densities into account, since clusters are of differing densities in different subspace cardinalities. To achieve this, we devise an innovative algorithm to adaptively determine the density thresholds for different cardinalities.

To extract clusters with different density thresholds in different cardinalities is useful but is quite challenging. Note that previous algorithms are infeasible in such an environment due to the lack of monotonicity property. That is, *if a k-dimensional unit is dense, any $(k-1)$-dimensional projection of this unit may not be dense.* Without the monotonicity property, the Apriori-like generate-and-test scheme adopted in almost previous works cannot be adopted for discovering all dense units in our model. A direct extension of previous methods is to execute a subspace algorithm once for each subspace cardinality k by setting the corresponding density threshold to find all k-dimensional dense units. However, it is very time consuming due to repeated execution of the targeted algorithm and repeated scans of database. To efficiently discover dense units, a practicable way would be to store the complete information of the dense units in all subspace cardinalities into a compact structure such that the mining process can be directly performed in memory without repeated database scans.

Motivated by this idea, we propose to construct a compact structure which is extended from the FP-tree [3] to

[1]For a cluster, recall is defined as the percentage of the data points in a true cluster, that are identified in this cluster. Precision is defined as the percentage of the data points in this cluster, that really belong to the true cluster.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011.

store the crucial information of the dataset. The base 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.

2. SUBSPACE CLUSTERING

We adopt the grid-based approach to discover subspace clusters, where the data space is partitioned into a number of non-overlapping rectangular units by dividing each attribute into δ equal-length intervals. Consider the projection of the data set in a k-dimensional subspace. A *"k-dimensional unit"* u is defined as the intersection of one interval from each of the k attributes. Let *count(u)* denote the number of data points contained in unit u. Note that the units in the same subspace cardinalities have the same size, and therefore we can use the count values to approximate the densities of the units in the same subspace cardinalities. Thus, the k-dimensional clusters can be discovered by first identifying the k-dimensional dense units, and then grouping the connected ones into clusters. Two k-dimensional units u_1, u_2 are connected if they have a common face or if there exists another k-dimensional unit u_3 such that both u_1 and u_2 are connected to u_3.

For identifying dense units, we propose to use different density thresholds for different subspace cardinalities. Let τ_k denote the density threshold for the subspace cardinality k, and let N be the total number of data points. In addition, an user input parameter α, called the *unit strength factor*, is introduced for specifying how dense a unit would be identified as a dense one. Then, we define the density threshold τ_k as:

$$\tau_k = \alpha \frac{N}{\delta^k}.$$

When the data are uniformly distributed in a k-dimensional subspace, the number of data points in each of the δ^k k-dimensional units in this subspace will be N/δ^k, i.e. the average unit density. In this scenario, there are no clusters discovered because everywhere in this space is almost of the same density. As the data are more compacted into clusters, the units within clusters will be much denser and would have a larger count value than the average density. Thus, the input parameter α is introduced such that a k-dimensional unit will be identified as a dense one if its count value exceeds α times of the average unit density, i.e., N/δ^k.

In addition, a user parameter, $k_{\max}$, is introduced for specifying the maximal subspace cardinality in such a way that clusters in cardinality up to $k_{\max}$ are discovered.

Problem Definition: Given the unit strength factor α and the maximal cardinality $k_{\max}$, for the subspaces with cardinality k from 1 to $k_{\max}$, find the clusters in which each is a maximal set of connected dense k-dimensional units whose count values exceed the density threshold τ_k.

3. APPROACH TAKEN

We explore a method to discover the dense units in subspace cardinality from 1 to $k_{\max}$. After the dense units are mined, we can follow the procedure proposed in [1] to discover the clusters, where the connected dense units are grouped into clusters. Therefore, we focuses on discovering the dense units in all subspaces.

The challenge of discovering dense units satisfying different density thresholds in different subspace cardinalities is that the monotonicity property no longer exists. That is, if a k-dimensional unit satisfies the threshold τ_k, any $(k-1)$-dimensional projection of this unit may not satisfy the threshold τ_{k-1}. Without the monotonicity property, the Apriori-like candidate generate-and-test scheme adopted in almost previous works cannot be adopted for discovering the dense units in our model.

To remedy this, we propose to first transform the dataset and then condense it with a compact structure for efficiently discovering dense units. The base 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. Note that a k-dimensional unit can be represented by the set of k 1-dimensional units, corresponding to the k intervals. By transforming each d-dimensional data record into d 1-dimensional units it resides in, the count value of a k-dimensional unit can be calculated by directly counting the occurrences of the set of k 1-dimensional units in the transformed dataset. Thus, finding k-dimensional dense units whose counts exceed the threshold τ_k is similar to mining frequent k-itemsets satisfying the minimum support. After the dataset is transformed, we can extend the FP-tree [3] for subspace clustering to store the complete information of the dense units satisfying different thresholds. Thus, dense units can be discovered efficiently from the tree, and then clusters are formed by grouping the connected dense units.

4. CONCLUSION

We in this paper studied the density conscious subspace clustering to take consideration of the densities into subspace clustering. An innovative algorithm was proposed to discover clusters in different cardinalities with different density thresholds such that clusters in different cardinalities can be discovered with both high precision and recall.

Acknowledgements

The work was supported in part by the National Science Council of Taiwan, R.O.C., under Contracts NSC93-2752-E-002-006-PAE.

5. REFERENCES

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *ACM SIGMOD*, 1998.

[2] C. H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based Subspace Clustering for Mining Numerical Data. *ACM SIGKDD*, 1999.

[3] J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. *ACM SIGMOD*, 2000.

[4] K. Kailing, H.-P. Kriegel, and P. Kroger. Density-Connected Subspace Clustering for High-Dimensional Data. *SIAM ICDM*, 2004.

[5] H. S. Nagesh, S. Goil, and A. Choudhary. Adaptive Grids for Clustering Massive Data Sets. *SIAM ICDM*, 2001.