Revealing True Subspace Clusters in High Dimensions

Jinze Liu, Karl Strohmaier, and Wei Wang
Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599
{liuj, strohma, weiwang}@cs.unc.edu

Abstract

Subspace clustering is one of the best approaches for discovering meaningful clusters in high dimensional space. One cluster in high dimensional space may be transcribed into multiple distinct maximal clusters by projecting onto different subspaces. A direct consequence of clustering independently in each subspace is an overwhelmingly large set of overlapping clusters which may be significantly similar. To reveal the true underlying clusters, we propose a similarity measurement of the overlapping clusters. We adopt the model of Gaussian tailed hyper-rectangles to capture the distribution of any subspace cluster. A set of experiments on a synthetic dataset demonstrates the effectiveness of our approach. Application to real gene expression data also reveals impressive meta-clusters expected by biologists.

Keywords: Subspace Clustering, Overlapping Cluster, Adhesion, Gaussian Tails, Cluster Intersection, Local Grid, Gene Expression

1 Introduction

Clustering has become one of the most popular and effective data mining techniques to reveal characteristics of large amounts of data. Because of the curse of the dimensionality, there have been many studies [2, 1, 3, 4] on designing new models and efficient algorithms for capturing subspace clusters.

Recent methods result in subspace clusters that are either *disjoint subspace clusters* or *overlapping subspace clusters*. In the model of disjoint clusters, each object may belong to at most a single cluster, regardless of the subspace in which the cluster is present. These disjoint clusters are succinct. In contrast, the model of overlapping clusters allows an object to naturally be included in multiple clusters, which offer indisputable advantages in real-world applications. For example, a gene may have multiple functions, each of which may be manifested in a specific metabolic pathway. It is biologically relevant to discover that one gene may appear in multiple gene clusters. However, the flexibility offered by this model poses substantial challenges in the design of the cluster model and mining algorithm. Due to efficiency concerns, existing models often discretize the data space and assume that all clusters are hyper-rectangles consisting of adjacent grids in their subspaces. Density and entropy measures based on data distribution within a subspace are applied to determine the existence of a subspace cluster. Since this decision is made independently for each subspace, clusters in different subspaces may overlap. In fact, the degree of cluster overlap may be very high, even if we only keep the set of maximal subspace clusters[1]. The number of subspace clusters can be large, which may degrade the significance of each cluster and hinder the posterior analysis. In this paper, we propose a new similarity measure for subspace clusters to accommodate the difference in subspaces, based on solid statistical theory. The application of our model to gene expression data revealed biologically relevant clusters with significantly reduced number of clusters.

Section 2 addresses related work in subspace clustering and cluster merging. Section 3 defines the model proposed in the paper. A greedy algorithm is presented in Section 4. The experiment is reported in Section 5. Section 6 concludes the paper and discusses some future work.

2 Related Work

Our work is related to grid and density-based subspace clustering, as well as cluster similarity analysis and cluster merging. Density- and grid-based approaches view clusters as regions of the data space in which objects are dense and are separated by regions of low object density. CLIQUE[2], MAFIA[4] and ENCLUS[3] are three such algorithms. The resolution of the grid ultimately determines the computation efficiency. These algorithms detect independent clusters in the highest dimensional subspaces, which may lead to a large number of clusters.

Previous work on cluster similarity assessment solely handles clusters in the full space. Classical hierarchical

[1] The formal definition of *maximal subspace cluster* is in Section 3. Intuitively, a maximum subspace cluster is a subspace cluster that cannot include more objects or dimensions without violating the clustering criteria.

clustering algorithms use the distance between clusters to measure similarity among clusters. In fuzzy clustering[5], the similarity of a pair of clusters is defined as the maximum percentage of data points of each cluster falling in the intersection of the two clusters. However, extra caution has to be used when merging clusters from different subspaces.

3 Model

Let $\mathcal{A} = \{A_1, A_2, ..., A_d\}$ be a set of ordered, numerical dimensions (attributes) with bounded domains $\mathcal{V}$ = $A_1 \times A_2 \times$ ...$\times A_d$ a d-dimensional vector space. We refer to each individual A_i ($i = 1, \ldots, d$) as a dimension (or attribute) of $\mathcal{V}$. The inner product of any subset of $\mathcal{A}$ forms a subspace of $\mathcal{V}$, which is denoted as $\mathcal{S}$. We assume that each dimension A_i has been normalized to range [0, 1). Let the database $\mathcal{O}$ be a set of data points (or objects) in $\mathcal{S}$. Each data point x is a d-dimensional unit vector.

We adopt the space discretization of CLIQUE[2] by superimposing a grid of non-overlapping rectangular units (cells) onto $\mathcal{S}$, given the size of the interval, λ, and the density threshold ξ. A *cluster* in a subspace is a maximum set of connected dense units. To discriminate from the cluster model we will propose later in the paper, we call it a *base cluster*. A base cluster is denoted as a tuple $C =< \mathcal{O}, \mathcal{S}, \mathcal{R} >$, where $\mathcal{O}(C)$ denotes the set of data points, $\mathcal{S}(C)$ denotes its subspace, and $\mathcal{R}(\mathcal{C})$ denotes the minimum bounding rectangle containing the cluster. A cluster within a k-dimensional subspace is called a k-dimensional subspace cluster. A cluster C is *maximum* if there does not exist another cluster C' such that $O(C) \subseteq O(C')$, and $S(C) \subseteq S(C')$. Given a subspace $\mathcal{S}$, let $\mathcal{CL}(\mathcal{S})$ be the set of base clusters generated by function $\mathcal{CL}$ in subspace $\mathcal{S}$.

3.1 Scenarios of Overlapping among Base Clusters

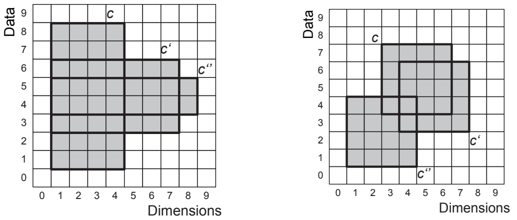
In our study, we classify the cluster overlap as *inclusive* overlap or *non-inclusive* overlap.

Lemma 3.1 *Given a maximal base cluster C in k-dimensional space $\mathcal{S}$, $\forall$ $(k-1)$-dimensional subspace $\mathcal{S}'$, $\mathcal{S}' \subset \mathcal{S}$, there exists $C' \in \mathcal{CL}(\mathcal{S}')$ such that $\mathcal{O}(C) \subseteq \mathcal{O}(C')$.*

Lemma 3.1 can be proven by the Apriori property of dense units in [2]. Given cluster C, the set of dense and connected units composing C are also dense and connected when they are projected onto any its $(k-1)$-dimensional subspace. Therefore, those dense and connected $(k-1)$-dimensional dense units may be equal to or part of a cluster C'. Lemma 3.1 describes the nested *inclusion* relationship occurring between clusters illustrated in Figure 1(a).

Corollary 3.1 *Given a k-dimensional space $\mathcal{S}$, let $\{\mathcal{S}'_1, \mathcal{S}'_2, \ldots, \mathcal{S}'_k\}$, $\mathcal{S}'_i \subset \mathcal{S}$, all be $k-1$-dimensional subspaces of $\mathcal{S}$. Given a cluster C, $C \in \mathcal{CL}(\mathcal{S})$, there exists $C'_i \in \mathcal{CL}(\mathcal{S}'_i)$ such that $\mathcal{O}(C) \subseteq \cap_{0<i\leq k}\mathcal{O}(C'_i)$.*

Beside inclusive overlap, base clusters residing in non-inclusive subspaces, where $\mathcal{S}' \subseteq \mathcal{S}$ and $\mathcal{S} \subseteq \mathcal{S}'$, may have non-inclusive overlap, where neither cluster is a subset of the other, as illustrated in Figure 1(b). Those overlapping clusters might be different views of one underlying cluster when they are projected onto corresponding subspaces.



(a)Inclusive overlap (b) Non-inclusive overlap

Figure 1. Two typical cases of cluster relationships.

In this study, all the base clusters are represented by minimum hyper-rectangular *kernels* defined by a pair of vectors $\overrightarrow{R^l}$ and $\overrightarrow{R^h}$, which are the lower and upper boundaries. The union of a set of base clusters is referred to as a *meta-cluster*. They are described by the hyper-rectangular kernels with softened Gaussian boundaries.

3.2 Adhesion Strength between Overlapping Clusters

Given two clusters, the *adhesion strength* is a measure of how tightly one cluster attaches to the other. Intuitively, two clusters have strong adhesion strength if both clusters have a large percentage of data points closely located. In addition, exclusive data points in the exclusive subspaces should not form statistically significant outliers.

Definition 3.1 *Given two clusters $C = \mathcal{O}(C) \times \mathcal{S}(C)$ and $C' = \mathcal{O}(C') \times \mathcal{S}(C')$, defined by two hyper-rectangular kernels R and R', the* **adhesion score** *of C to C' is defined as*

$$\mathcal{H}(C, C') = \sqrt[|S(C)|]{\prod_{i \in \mathcal{S}(C)} h(C, C', i)} \qquad (1)$$

where the adhesion score along each dimension i is defined as

$$h(C, C', i) = \frac{|\overrightarrow{R_i} \wedge \overrightarrow{R'_i}|}{|\overrightarrow{R_i}|} \frac{\mathcal{Q}(\overrightarrow{R_i} \wedge \overrightarrow{R'_i}, C)}{\mathcal{Q}(\overrightarrow{R_i}, C)} \qquad (2)$$

where function $\mathcal{Q}(\overrightarrow{R}, C)$ *returns the number of points in* C *that falls in region* $\overrightarrow{R}$*, and* $\overrightarrow{R_i} \wedge \overrightarrow{R'_i}$ *is*

$$
\begin{aligned}
&(i).\ 0,\\
&\qquad \textbf{if } (i \in \mathcal{S}(C) \cap \mathcal{S}(C'))\\
&\qquad \textbf{and } (max(\overrightarrow{R_i^l}, \overrightarrow{R'^l_i}) > min(\overrightarrow{R_i^h}, \overrightarrow{R'^h_i})),\\
&\qquad\qquad (Figure\ 2(b)(1));\\
&(ii).\ [min(\overrightarrow{R_i^h}, \overrightarrow{R'^h_i}), max(\overrightarrow{R_i^l}, \overrightarrow{R'^l_i})],\\
&\qquad \textbf{if } (i \in \mathcal{S}(C) \cap \mathcal{S}(C'))\\
&\qquad \textbf{and } (min(\overrightarrow{R_i^h}, \overrightarrow{R'^h_i}) > max(\overrightarrow{R_i^l}, \overrightarrow{R'^l_i})),\\
&\qquad\qquad (Figure\ 2(b)(2.1\&2.2));\\
&(iii).\ [\min_{x \in \mathcal{O}(C')} x_i, \max_{x \in \mathcal{O}(C')} x_i] \wedge \overrightarrow{R_i}\\
&\qquad \textbf{if } (i \in \mathcal{S}(C)) \textbf{ and } (i \notin \mathcal{S}(C')).
\end{aligned}
\tag{3}
$$

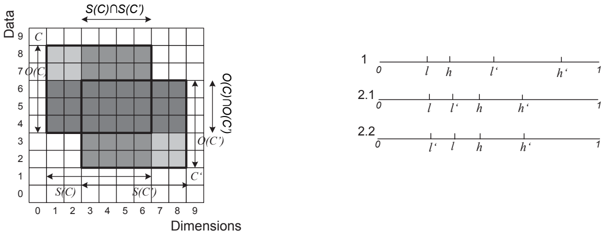
The adhesion strength is asymmetric and can be used to construct a similarity measure between two clusters.

$$Similarity(C, C') = \min\{\mathcal{H}(C, C'), \mathcal{H}(C', C)\}.$$

In Equation 2, the adhesion strength from one cluster to the other in each dimension depends on the ratio between the intersection area and the cluster boundary, as well as the percentage of data points falling in the intersection region. In each dimension, the adhesion strength h of two clusters is captured in terms of both data points and physical space. Two clusters can adhere to each other only if they have a common kernel(Figure 2(b) 2.1&2.2), which is the intersection of their hyper-rectangular kernels. Furthermore, both of the clusters should have a large percentage of data points falling within the kernel. This is determined by function $\mathcal{Q}$. If two clusters do not have a boundary intersection in any dimension of their subspaces, the adhesion strength is 0 (Figure 2 (b) 1). In the exclusive dimensions, when only the cluster who owns the dimension has a kernel boundary, we approximate the other kernel boundary using the maximum and minimum of the common points in that dimension (Case (iii) in Definition 3.1). The overall adhesion strength across all unified dimensions is the geometric average of the adhesion strength in each dimension. The asymmetric property of adhesion strength enables its use in describing the scenario when one cluster is much smaller than the other but is strongly coupled to it. If the similarity between a pair of clusters is high, we may merge them into a new meta-cluster. If C and C' are the two meta-clusters of high similarity, the new meta-cluster C'' will be in the subspace $\mathcal{S}(C) \cup \mathcal{S}(C')$ and include objects $\mathcal{O}(C) \cup \mathcal{O}(C')$. To determine the hyper-rectangular kernel of the new meta-cluster, a Gaussian tailed rectangular kernel is fit onto the cluster based on the model proposed in [8].

3.2.1 Rectangular Kernel Determination

Let C_1 and C_2 be the two clusters with kernels $[\overrightarrow{R^{l_1}}, \overrightarrow{R^{h_1}}]$ and $[\overrightarrow{R^{l_2}}, \overrightarrow{R^{h_2}}]$. They are to be merged to generate a meta-cluster C.



(a)Two overlapping clusters. (b) Two intervals relationships.

Figure 2. Illustration of Definition 3.1.

$$
\begin{aligned}
LL(x) = &(-|O_1 \cup O_2| \ln(\sqrt{2\pi}\sigma_i + \overrightarrow{R_i^{h_3}} - \overrightarrow{R_i^{l_3}}))\\
&+ \textstyle\sum_{x \in O_1 \cup O_2} -\frac{1}{2}\left(\frac{x_i - closest(x_i, \overrightarrow{R_i^{l_3}}, \overrightarrow{R_i^{h_3}})}{\sigma_i}\right)
\end{aligned}
\tag{4}
$$

The optimal values of $\overrightarrow{R_i^{l_3}}$ and $\overrightarrow{R_i^{h_3}}$ should be the ones that maximize $LL(x)$ in Equation 4. The golden ratio optimizer [7] can be used to locate the optimal values of $\overrightarrow{R_i^{l_3}}$ and $\overrightarrow{R_i^{h_3}}$, starting with the initial values $[\overrightarrow{R_i^{l_3}}, \overrightarrow{R_i^{h_3}}] = [\overrightarrow{R_i^{l_1}}, \overrightarrow{R_i^{h_1}}] \wedge [\overrightarrow{R_i^{l_2}}, \overrightarrow{R_i^{l_2}}]$, as defined in Definition 3.1. The intersection area for measuring adhesion strength is set as the initial kernel to meet its density requirement. Once we determine the initial kernel of the meta-cluster, we may apply the MLE of the Gaussian tailed hyper-rectangular distribution to compute its optimal kernel.

4 A Greedy Algorithm

This algorithm evaluates every pair of clusters at each iteration, and picks the pair having the best adhesion score and merges it. After that, the adhesion score of those pairs having one of the clusters being merged is updated with the new meta-cluster and reordered. If there still exists a pair of clusters having an adhesion score above the threshold, the merging continues. This algorithm is simple and straightforward. The time complexity is $O(N^3 logN)$, where N is the total number of clusters.

```
Algorithm MergeGreedy($C, \delta$)
Input: $C$: A set of unique dense subspace clusters; $\delta$: similarity threshold
Output: $C$': A set of unique subspace clusters
Compute the similarity score of $\frac{|C|*(|C|-1)}{2}$ cluster pairs
Sort the list of cluster pairs in decreasing order of the score
while Head of the list having a similarity score> $\delta$
    do Merge $C_i$, $C_j$ with highest score.
        Remove cluster pairs containing $C_i$ or $C_j$
        Recompute the similarity score.
        Insert them into the list of cluster pairs in decreasing order of the score.
return.
```

5 Performance Evaluation

We compare the meta-clusters with the base clusters generated by CLIQUE in terms of the cluster quality using both synthetic data sets and real gene expression data. All implementations are in C++ and tested on a machine with an 800MHz Pentium III processor and 2GB of main memory.

5.1 Effect of Meta-Clustering

The synthetic high dimensional data set is generated by embedding clusters in the subspaces. The clusters are points following a Gaussian tailed hyper-rectangular distribution. For each cluster, the number of data points and the hyper-rectangular kernel are first determined by selecting the dimensionality of the subspace, and the upper and lower bounds of the hyper-rectangular kernel. Let p be the percentage of data points we want to put in the kernel. We have $\sigma = \frac{(1-p)(\mathcal{R}^h - \mathcal{R}^l)}{\sqrt{2\pi}}$. A random generator is used to distribute data points into the kernel, $\pm\sigma$ outside the kernel and $\pm 2\sigma$ outside the kernel for each dimension, accordingly.

In this group of experiments, we embedded 5 6-dimensional meta-clusters in the data set with 1500 20-dimensional data points. The experiment varies the adhesion score from 0.4 to 1 in steps of 0.1. The number of clusters in each dimension is presented in Figure 3. The clusters generated when the adhesion score is 1 are actually the base clusters, which corresponds to the top curve in the figure. The general trend is that fewer meta-clusters are generated with a lower adhesion score because more clusters may be valid for merging. The big gap between the curves of adhesion scores 0.6 and 0.7 suggests that 0.6 to 0.7 may be a good point to have a reasonable number of meta-clusters matching well with the underlying real clusters.

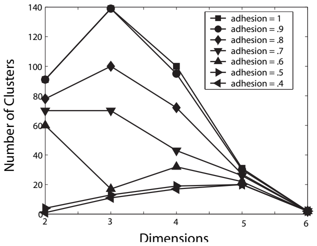


Figure 3. Number of clusters in different subspaces with varying adhesion score.

5.2 Meta-Clustering on real data

Two types of cell lines derived from basal epithelium and luminal epithelium respectively were treated under the chemotherapeutics. The expression levels of genes of both cell lines are recorded during the treatment at 12, 24, and 36 hours. Multiple samples are generated at each time point. After certain filtering of noise, we selected 1034 genes and 26 columns for analysis. Each type of cell line has 13 columns in the gene expression matrix. On the advice of biologists, we normalized the expression levels using logarithms and then transformed them into [-1, 1]. We divided each dimension into 5 intervals with length 0.4. The density threshold is 0.01. The whole clustering process took less than two minutes, and given an adhesion threshold of 0.6, it generated 28086 base clusters, from which 4130 meta clusters were generated. That is only 14% of the number of base clusters. This number is reasonable considering the very low density threshold and the number of subspaces.

6 Conclusion

In this paper, we provide a framework to organize the overlapping subspace clusters generated in grid and density-based algorithms. We show that significant overlap among clusters is very common in subspace clustering and can result in redundancy in identifying real clusters embedded in a data space. Adhesion strength is defined to measure the similarity between two clusters with a Gaussian tailed hyper-rectangular shape. Experiments on both synthetic and real datasets highlight the effectiveness of the adhesion strength in measuring the similarity of subspace clusters. Our ongoing work includes developing simple and efficient algorithms for meta-clustering.

Acknowledgement

We thank Dr. Chuck Perou and his lab at UNC for their valuable help with gene expression experiments.

References

[1] C. Aggarwal, C. Procopiuc, J.L.Wolf, P.S.Yu and J.S.Park. A Framework for Finding Projected Clusters in high dimensional spaces. In SIGMOD, 1999.

[2] R.Agrawal, J.Gehrke, D.Gunopulos, and P.Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD, 1998.

[3] C.H. Cheng, A.W. Fu, Y. Zhang, Entropy-based Subspace Clustering for Mining Numerical Data. In SIGKDD, Aug 1999.

[4] S. Goil, Harsha and Alok Choudhary. MAFIA: Efficient and scalable subspace clustering for very large data sets. In ICDE, 2000.

[5] U.Kaymak, and M.Setnes. Extended Fuzzy Clustering Algorithms. Discussion paper of Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam, 2000.

[6] J. Liu, K. Strohmaier and W. Wang. Revealing Subspace Clusters in High Dimensional Space. UNC-CH CS technical report.

[7] W.H.Press, S.A. Teukolsky, W.T.Vetterling, B.P.Flannery. Numerical recipes in C, 2_{nd} edition.

[8] Dan Pelleg, Andrew Moore, Mixtures of Rectangles:Interpretable Soft Clustering, ICML, 2001