A REVIEW ON SPARSE SUBSPACE CLUSTERING FOR HIGH DIMENSIONAL DATA (HDD)

Ambika P.R 1, Dr. Bharathi Malakreddy A 2
1Department of Computer Science & Engineering, City Engineering College, Bengaluru
2Department of Information Science & Engineering, BMSIT&M, Bengaluru

ABSTRACT:

Sparse Subspace Clustering constructs a sparse similarity graph by using the coefficient of sparse representation to subspace clustering. Subspace clustering has been investigated extensively since traditional clustering algorithms often fail to detect meaningful clusters in high-dimensional data spaces. Often these high-dimensional data can be represented by a low-dimensional subspace. In this paper, we present several state-of-art techniques for analyzing high dimensional data, e.g., frequent pattern mining, clustering, and classification. We will discuss how these methods deal with the challenges of high dimensionality. Based on a survey on subspace clustering, we identify challenges and issues involved with clustering HDD.

Keywords: High Dimensional, HDD, Sparse Space Clustering, Dimensionality Reduction, Curse of dimensionality.

[1] INTRODUCTION

Many real-world problems on computer vision and image processing applications require processing and representation of high-dimensional data such as images, videos, text and web documents, DNA microarray data, and more. Often, such high-dimensional data lie close to low-dimensional structures corresponding to several classes or categories to which the data belong. Subspace clustering has been investigated extensively since traditional clustering
algorithms often fail to detect meaningful clusters in high-dimensional data spaces. Many recently proposed subspace clustering methods suffer from two severe problems: First, the algorithms typically scale exponentially with the data dimensionality and/or the subspace dimensionality of the clusters. Second, for performance reasons, many algorithms use a global density threshold for clustering, which is quite questionable since clusters in subspaces of significantly different dimensionality will most likely exhibit significantly varying densities.

In this paper, we study an algorithm, called Sparse Subspace Clustering (SSC), to cluster data points that lie in a union of low dimensional subspaces. The key idea is that, among the infinitely many possible representations of a data point in terms of other points, a sparse representation corresponds to selecting a few points from the same subspace as shown in figure 1. This motivates solving a sparse optimization program whose solution is used in a spectral clustering framework to infer the clustering of the data into subspaces. When clustering high dimensional data, traditional clustering methods are found to be lacking since they consider all of the dimensions of the dataset in discovering clusters whereas only some of the dimensions are relevant. This may give rise to subspaces within the dataset where clusters may be found. Using feature selection, we can remove irrelevant and redundant dimensions by analyzing the entire dataset. The problem of automatically identifying clusters that exist in multiple and maybe overlapping subspaces of high dimensional data, allowing better clustering of the data points, is known as subspace clustering. There are two major approaches to subspace clustering based on search strategy. Top-down algorithms find an initial clustering in the full set of dimensions and evaluate the subspaces of each cluster, iteratively improving the results. Bottom-up approaches start from finding low dimensional dense regions, and then use them to form clusters. Based on a survey on subspace clustering, we identify the challenges and issues involved with clustering of HDD.

![Figure 1. Sparse subspace clustering](image)

[2] RELATED WORK

Various algorithms have been proposed in the literature for subspace clustering. Some of these algorithms are iterative in nature while others are statistical and algebraic. Approaches based on spectral clustering have also been proposed in the literature. In particular, sparse representation and low-rank approximation-based methods for subspace clustering have gained a lot of traction in recent years. These methods find a sparse or low-rank representation of the data and build a similarity graph whose weights depend on the sparse or low-rank coefficient matrix for segmenting the data. One of the advantages of these methods is that they are robust to noise and occlusion. Furthermore, some of these approaches do not require the knowledge of the dimensions and the number of subspaces. In particular, the Sparse Subspace Clustering (SSC) algorithm, Low-Rank Subspace Clustering (LRSC) and Low-Rank Sparse Subspace Clustering (LRSSC) method are well supported by theoretical analysis and provide state-of-the-
art results on many publicly available data sets such as the Hopkins155 benchmark motion segmentation dataset. Finding sparse or low-rank representation is very computationally demanding especially when the dimension of the features is high. This is one of the drawbacks of the sparse and low-rank representation-based methods. To deal with this problem, dimensionality reduction is generally applied on the data prior to applying these algorithms. Sub-space clustering algorithms are divided into two categories (a) Bottom-Up Search (b) Top-Down Search.

(a) Bottom-Up Subspace Search Methods

This category of method take advantage of the downward closure property of density to reduce the search space using an APRIORI style approach. This algorithm's first step is to create a histogram for each dimension and selecting those bins with densities above a given threshold. The algorithms proceeds until there are no more dense units found. Adjacent dense units are grouped to form clusters. Some algorithms based on these approach are; CLIQUE, ENCLUS, MAFIA, CLTree, DOC and Cell Based Clustering(CBF). The advantages of Bottom-Up approach are; i) able to handle High Dimensional Data ii) can find clusters of arbitrary sizes and shapes scale reasonably with the amount of data. These method also has some disadvantages ; i) the algorithms do not scale well with the increase in number of dimensions ii) may eliminate small clusters iii) the run time increases exponentially with the increase in the number of dimensions in the datasets.

(b) Top-Down Iterative search methods

These approach starts by finding an initial approximation of the clusters in the full feature space with equally weighted dimensions.

In particular, the Sparse Subspace Clustering (SSC) algorithm, Low-Rank Subspace Clustering(LRSC) and Low-Rank Sparse Subspace Clustering (LRSSC) method are well supported by theoretical analysis and provide state-of-the-art results on many publicly available data sets such as the Hopkins155 benchmark motion segmentation dataset. Finding sparse or low-rank representation is very computationally demanding especially when the dimension of the features is high. This is one of the drawbacks of the sparse and low-rank representation-based methods. To deal with this problem, dimensionality reduction is generally applied on the data prior to applying these algorithms. Dimensionality reduction methods such as Principle Component Analysis(PCA) and Random Projections (RP) can reduce the dimension of data. However, a well learned projection matrix can lead to a higher clustering accuracy at a lower dimensionality. Several works have been proposed in the literature that find a sparse representation on a low-dimensional latent space. However, these methods are specifically designed for classification tasks and not for clustering. Motivated by some of the sparsity promoting dimensionality reduction methods, in this paper, we propose methods for simultaneous dimensionality reduction and subspace clustering under the frameworks of SSC,LRR and LRSSC.

[3] High Dimensional Data-HDD

High Dimensional Data(HDD) is the new emerging data which has a large amount of attributes. Typical examples of high-dimensional data in which task-specific methods have been proposed include: gene expression analysis, text document analysis, video analysis and customer recommendation systems. Challenges faced by high-dimensional data are different from those posed by low-dimensional data and therefore potentially require different approaches to addressing them. Clustering high-dimensional data has been a major challenge due to the inherent sparsity of the points. Most existing clustering algorithms become substantially inefficient if the required similarity measure is computed between data points in the full dimensional space. High dimensional data is increasingly common in many fields. As the number of dimensions increase, many clustering techniques begin to suffer from the curse of dimensionality, degrading the quality of the results. In high dimensions, data becomes very sparse and distance measures become increasingly meaningless. This problem has been studied
extensively and there are various solutions, each appropriate for different types of high dimensional data and data mining procedures. Figure 2. shows that clusters are embedded in different subspaces of high-dimensional datasets.

Figure 2. An Example of subspace clusters.

**Curse of Dimensionality**

The ‘curse of dimensionality’ states that as the number of dimensions increases, the concept of distance as measured between points become less and less meaningful. Common challenges for analyzing HDD are: The first one is the curse of dimensionality. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications. Secondly, the specificity of similarities between points in a high dimensional space diminishes. It was proven in [3] that, for any point in a high dimensional space, the expected gap between the Euclidean distance to the closest neighbor and that to the farthest point shrinks as the dimensionality grows. This phenomenon may render many data mining tasks (e.g., clustering) ineffective and fragile because the model becomes vulnerable to the presence of noise.

A traditional and common method to deal with high dimensionality is dimensionality reduction in the form of feature selection or feature transformation. A feature selection or feature transformation method reduces the original feature space so that the data set is clustered more accurately under the new reduced feature space. Feature selection approaches find a single subset of attributes to group all data points, which can be different than reality for a lot of applications. On the other hand, feature transformation finds linear combinations of attributes from the original feature space and replaces a subset of features with a single attribute representing the feature subset. The resulting reduced feature space might improve clustering performance, but the solution does not convey any further useful information once the original feature space is lost. Also similar to feature selection methods, feature transformation approaches find a single subspace for all data set instances thus failing to identify subspace structure when it exists.

**Dimensionality Reduction**

High dimensional data occur naturally in many domains and have presented great challenges for traditional data mining techniques. Traditional clustering algorithms become computational expensive when data set to be cluster is large. The curse of dimensionality refers to various phenomena that arise when analyzing data is high dimensional data that do not occur in low dimensional data. The common theme of this problem is that, when the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse. So that in high dimensional data all objects appear to be dissimilar in many ways so it becomes difficult to cluster.

[4] **SPARSE SUBSPACE CLUSTERING (SSC)**

In this section, we introduce the Sparse Subspace Clustering algorithm for clustering a collection of multi-subspace data using sparse representation techniques. The underlying idea
behind the algorithm is the self-expressiveness property of the data, which represents that each data point in a union of subspaces can be efficiently represented as a linear or affine combination of other points. Such a representation is not unique in general because there are infinitely many ways in which a data point can be expressed as a combination of other points. The key observation is that a sparse representation of a data point ideally corresponds to a combination of a few points from its own subspace. The goal of SSC is to identify the number of subspaces, their dimensions, a basis for each subspace and membership of each data point to its correct subspace.

Problem Statement

Given a set of data points lying in a union of unknown linear subspaces, there are n subspaces $s_1,s_2,...,s_n$ of $\mathbb{R}^D$ of dimensions $d_1,d_2,...,d_n$. Assume we are given a set of N data points $y_1,y_2,....,y_n$, the aim of subspace segmentation is to approximate the underlying subspaces by using the set of data points.

The Steps of SSC Method

SSC is based on the theory of sparse representation, which takes into account of the self-expressions property of the data points. Under some conditions, SSC method is effective for handling subspace segmentation problems. The steps of SSC are as follows:

Input: Assume we are given a set of N data points $y_1,y_2,....,y_n$, lying near a union of n subspaces $s_1,s_2,...,s_n$ of $\mathbb{R}^D$ of dimensions $d_1,d_2,...,d_n$.

1. Solve the sparse representation optimization problem
2. Construct a similarity graph G with nodes representing the N data points
3. Sort the eigenvalues of the normalized Laplacian matrix of the graph G in descending order.
4. Apply the spectral clustering to the similarity graph in order to obtain the partition $y_1,y_2,....,y_n$.

Output: Obtain the segmentation of the data points: $y_1,y_2,....,y_n$.

Clustering High Dimensional Data

The objects in data mining could have hundreds of attributes. Clustering in such high dimensional spaces presents tremendous difficulty, much more so than in predictive learning. In decision trees, for example, irrelevant attributes simply will not be picked for node splitting, and it is known that they do not affect Naïve Bayes as well. In clustering, however, high dimensionality presents a dual problem. First, under whatever definition of similarity, the presence of irrelevant attributes eliminates any hope on clustering tendency. After all, searching for clusters where there are no clusters is a hopeless enterprise. While this could also happen with low dimensional data, the likelihood of presence and number of irrelevant attributes grows with dimension. The second problem is the dimensionality curse that is a loose way of speaking about a lack of data separation in high dimensional space. Mathematically, nearest neighbor query becomes unstable: the distance to the nearest neighbor becomes indistinguishable from the distance to the majority of points [Beyer et al. 1999]. This effect starts to be severe for dimensions greater than 15. Therefore, construction of clusters founded on the concept of proximity is doubtful in such situations. For interesting insights into complications of high dimensional data, see [Aggarwal et al. 2000]. Basic exploratory data analysis (attribute selection) preceding the clustering step is the best way to address the first problem of irrelevant attributes. Idea to group items very similar to co-clustering has already been discussed in the section Co-Occurrence of Categorical Data.
APPLICATIONS
Limitations of current methods and the application of subspace clustering techniques to new domains drives the creation of new techniques and methods. Subspace clustering is especially effective in domains where one can expect to find relationships across a variety of perspectives. Some areas where we feel subspace clustering has great potential are information integration system, text-mining, and bioinformatics. Subspace clustering techniques can be leveraged to uncover the complex relationships found in data from each of these areas.

1. Information Integration Systems
Information integration systems are motivated by the fact that our information needs of the future will not be satisfied by closed, centralized databases. Rather, increasingly sophisticated queries will require access to heterogeneous, autonomous information sources located in a distributed manner and accessible through the Internet. In this scenario, query optimization becomes a complex problem since the data is not centralized. The decentralization of data poses a difficult challenge for information integration systems, mainly in the determination of the best subset of sources to use for a given user query.

2. Web Text Mining
The explosion of the world wide web has prompted a surge of research attempting to deal with the heterogeneous and unstructured nature of the web. A fundamental problem with organizing web sources is that web pages are not machine readable, meaning their contents only convey semantic meaning to a human user. In addition, semantic heterogeneity is a major challenge. That is when a keyword in one domain holds a different meaning in another domain making information sharing and interoperability between heterogeneous systems difficult.

3. DNA Microarray Analysis
DNA microarrays are an exciting new technology with the potential to increase our understanding of complex cellular mechanisms. Microarray datasets provide information on the expression levels of thousands of genes under hundreds of conditions. For example, we can interpret a lymphoma dataset as 100 cancer profiles with 4000 features where each feature is the expression level of a specific gene. This view allows us to uncover various cancer subtypes based upon relationships between gene expression profiles. Understanding the differences between cancer subtypes on a genetic level is crucial to understanding which types of treatments are most likely to be effective.

[5] CONCLUSION
Subspace clustering has been investigated extensively since from traditional clustering approaches often fail to detect meaningful clusters in high dimensional data spaces. Over the past few decades, significant progress has been made in clustering high-dimensional data sets distributed around a collection of linear and affine subspaces. This paper presented a review of such progress which included a number of existing subspace clustering algorithms, which were designed under the assumptions of perfect data and knowledge of the number of subspaces and their dimensions, throughout the years algorithms started to handle noise, outliers, data with missing entries, unknown number of subspaces and unknown dimensions.

Future Directions
In the case of noisy data, the theoretical correctness of existing algorithms is largely untouched. In the case of noiseless data retrieved from linear subspaces, the theoretical correctness of existing algorithms is well studied. Some subspace algorithms are correct for independent subspaces, others are correct for independent subspaces, others are provably correct for disjoint subspaces and others are able to handle an unknown number of subspaces of unknown dimensions in an arbitrary configuration. However, the development of theoretically sound algorithms for finding the number of subspaces and their dimension in the presence of noise and outliers is a very important open challenge. In our view, the grand challenge for the next decade will be to develop clustering algorithms for data drawn from multiple nonlinear
manifolds. Our interest in subspace clustering stems from the need to cluster text, which is often represented as a high-dimensional, sparse data set in the bag of words model, and where the number of attributes is in the thousands. Finally, we believe that the approach of identifying a small number of relevant attribute groups opens up possibilities for visualizing the second stage clustering problem, which allows applying interactive data clustering techniques.

REFERENCES

[1] Ehsan Elhamifar Rene Vidal, "Sparse Subspace Clustering" Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA.


