

# Multi-view Subspace Clustering for High-dimensional Data

Xiaojun Chen and Joshua Zhexue Huang  
Shenzhen Institutes of Advanced Technology,  
Chinese Academy of Sciences,  
Shenzhen 518055, China.

## Extended Abstract

The data today is towards more observations and very high dimensions. Large high-dimensional data are usually sparse and contain many classes/clusters. For example, large text data in the vector space model often contains many classes of documents represented in thousands of terms. It has become a rule rather than the exception that clusters in high-dimensional data occur in subspaces of data, so subspace clustering methods are required in high-dimensional data clustering.

Many subspace clustering algorithms have been proposed to handle high-dimensional data, aiming at finding clusters from subspaces of data, instead of the entire data space. They can be classified into two categories: hard subspace clustering that determines the exact subspaces where the clusters are found, and soft subspace clustering that assigns weights to different features and discover clusters from the subspaces of the features with large weights.

Many high-dimensional data sets are the results of integration of measurements on observations from different perspectives so that the features of different measurements can be grouped. For example, the features of the nucleated blood cell data were divided into groups of density, geometry, “color” and texture, each representing one set of particular measurements on the nucleated blood cells. In a banking customer data set, features can be divided into a demographic group representing demographic information of customers, an account group showing the information about customer accounts, and the spending group describing customer spending behaviors. Web pages can be represented with three views: a term vector view whose elements correspond to the occurrence of certain words in the web page text, a hyperlink graph view that shows other web pages which each web page points to, and a term vector view for the words in the anchor text. Such data is called **multi-view data**, which contains multiple views (representations/variable groups) from different feature spaces.

In the **multi-view data**, the objects in these data sets are categorized jointly by multiple views but the importance of different views varies in different clusters. The view level difference of features represents important information to subspace clusters and should be considered in the subspace clustering process. This is particularly important in clustering high-dimensional data because the weights of individual features tend to be less different when the number of the features becomes very large. However, the existing subspace clustering algorithms fail to make use of multiple views information in clustering high-dimensional data.

In the past decade, multi-view data has raised interests in the so-called multi-view clustering. Different from the traditional clustering methods which take multiple views as a flat set of variables and ignore the difference among different views, multi-view clustering exploits the information from multiple views in order

to produce a more accurate and robust partitioning of the data.

In this talk, we present a new multi-view subspace clustering method for clustering high-dimensional data in subspaces at the view level and individual feature level. In this method, a weight is assigned to each view in each cluster to identify the importance of the view in categorizing the cluster. In the meanwhile, a weight is also assigned to each feature in each cluster to identify the importance of the feature in categorizing the cluster. We present the results of a series of experiments on both simulation and real life data sets to show that the new subspace algorithm improved clustering performance significantly in comparison with other clustering algorithms, including the standard k-means, W-k-means, LAC and EWKM.

**Keywords:** Data Mining, Subspace Clustering, Multi-view Clustering