Abstract

Clustering is a fundamental task in data analysis, which partitions a set of samples into homogeneous groups. Over the past few years, graph-based clustering algorithms have become very popular and widely used in various applications due to the high efficiency and simplicity in implementation, i.e., graph-based clustering only requires a typical pairwise similarity matrix of samples (a.k.a. a graph) as input and performs spectral decomposition on that matrix to generate the clustering result. Despite its popularity, some fundamental issues still exist, which greatly affect its performance.

First, the performance of graph-based clustering algorithms is highly dependent on the quality of the input similarity matrix. Therefore, how to construct a high-quality graph which can best capture the underlying structures of data is the core of this kind of method. Besides, graph learning is also a hot research topic in signal processing. However, it is a challenging task. The traditional methods usually use the Euclidean distance of samples to construct the pairwise similarity matrix, which is very sensitive to the noise and usually suffers from high computational complexity. Besides, in practice, there is some side information available that may provide valuable information to promote graph-based clustering, such as the available supervisory information, which may be overlooked or not fully exploited to some extent. We will propose a unified theoretical framework capable of fully exploring the available side information, including the raw features and some week supervisory information, to learn a more robust and informative similarity matrix for graph-based clustering.

Second, most of the traditional graph-based clustering methods adopt linear models to cluster the data, but in practice, the data is not necessarily located in the linear subspaces, which may significantly affect the performance of graph-based clustering algorithms. It becomes insufficient and unreliable to only use the linear models in graph-based clustering. In addition, inspired by the powerful representation capability of the deep learning framework to model the nonlinear system, in this project, we will develop a novel deep neural network architecture to learn a nonlinear mapping of the data to well adapt to the subspace clustering.

With our solid backgrounds and the promising verifications achieved, it is highly expected that our investigations will provide more efficient and effective graph-based clustering algorithms for the large-scale dataset. Clustering is a fundamental and universal problem related to many topics, including low-rank matrix/tensor approximation, low-dimensional embedding, semi-supervised learning, noise estimation, deep learning, etc. In addition, the constructed high-quality graph can be widely used in various tasks in data analysis rather than just limited to clustering. We believe that beyond the three years envisioned for this work, the scientific findings of this project will continuously motivate the research on a wide range of research communities and benefit various real-world applications.