Sparse subspace clustering for large scale hyperspectral data

Problem statement

Supervised classification methods such as the classical support vector machine (SVM) and the modern convolutional neural network (CNN) require labeled training samples to train the classification model. Since data labeling is typically labor intensive and time-consuming, labeled samples are not always available in practice, posing serious limitations for supervised classification. Clustering, as an unsupervised approach, is able to effectively discriminate data points belonging to different clusters without any labeled data. Thus, clustering has much wider applications than supervised classification, especially in the dynamic scenarios like monitoring forest fires, disaster damages, land use/cover change detection and trajectory data mining where clustering methods play the unique role.

We focus on the subspace clustering technique that yields state-of-the-art clustering performance in the fields of computer vision, image processing, remote sensing and pattern recognition. The main idea is to model the input data by a union of subspaces and uncover the cluster structure in lower-dimensional subspaces, as shown in Fig. 1. Compared with the classical fuzzy c-means and k-means methods, subspace clustering approaches are able to unveil more precisely the data correlations, leading to superior clustering performance. We are particularly interested in the processing of hyperspectral images (HSIs) in remote sensing, where the goal in this proposal is to cluster pixels of HSI into different groups by using their spectral signatures, as shown in Fig. 2.

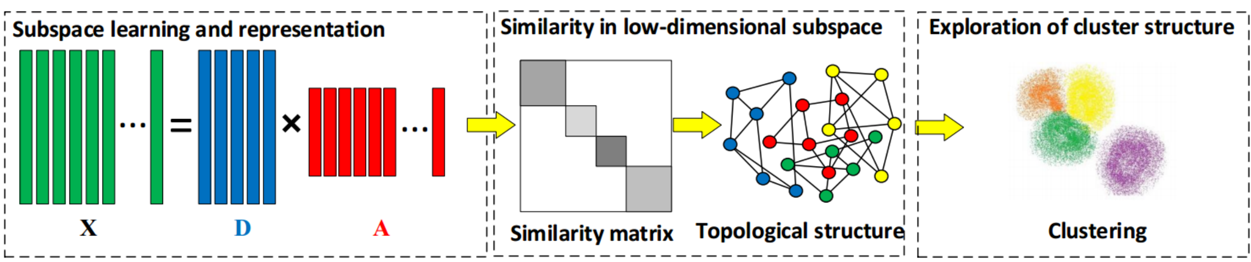

Fig. 1. The framework of a typical subspace clustering method, which includes subspace learning and representation, graph construction and spectral clustering. $\mathbf{X}$ is an input matrix with each column representing a data point; $\mathbf{D}$ is a dictionary that models the underlying subspaces; $\mathbf{A}$ is the corresponding subspace representation matrix with respect to $\mathbf{D}$.

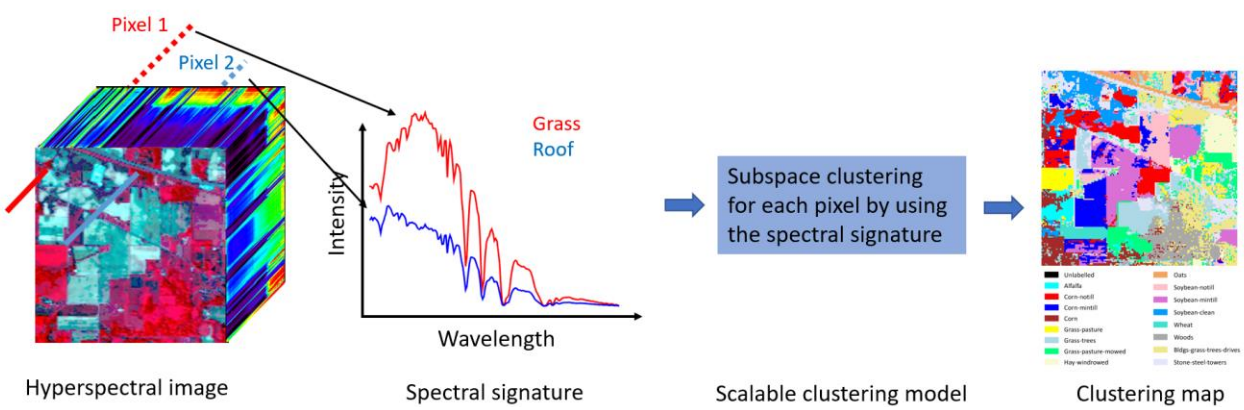

Fig. 2. An illustration of subspace clustering in the application of hyperspectral remote sensing images.

Despite the excellent clustering accuracy of subspace clustering techniques, their high computational complexity limits real applications involving big data sets, especially in real-time processing tasks. It is therefore important to reduce the computational complexity and develop scalable subspace clustering methods for large-scale data, which would facilitate their practical application. Important aspects in addressing this problem are understanding representation learning (including dictionary learning and subspace representation) and efficient algorithm design. Research group GAIM has rich experience in this domain and will provide full support in programming, model construction, optimization algorithm and experiment validation based on the well-founded expertise.

Goal:

The goal of this Master thesis is to advance further the current subspace clustering methods in terms of computational complexity and clustering accuracy, making a scalable and fast clustering algorithm accessible to the large-scale hyperspectral data. The contents of this research will include:

- Designing or learning a compact and discriminative dictionary to model the underlying low-dimensional subspaces of data.
- Extending the current subspace clustering method with the developed dictionary such that it has a lower computational complexity and comparable or even improved accuracy and thus is appliable for large-scale hyperspectral data.
- Developing an efficient algorithm to solve the resulting model and conducting experiments to validate the effectiveness of the proposed approach in real hyperspectral remote sensing images.

The students will start with the current subspace clustering code of the best available GAIM's technique and will also have other useful GAIM's techniques in terms of dictionary learning, sparse coding and optimization algorithm.