Improved Graph Based Clustering Methods and its Applications

Fahad Mostafa, Advisor: Dr. Victoria Howle
Department of Mathematics and Statistics, Texas Tech University

Clustering algorithms are frequently used on large data sets to analyze and find key patterns and rules in large data sets such as image data, gene expression data, or omics data sets. Graph clustering divides comparable objects into clusters based on graph properties in the data; K-means clustering is a distance-based or centroid-based clustering algorithm; and spectral clustering identifies strongly related communities on a network, using a normalized Laplacian matrix determined from the data's graph structure. To find the first *p* eigenvectors of a normalized Laplacian matrix, power iteration can be straightforward, quick, and reasonably scalable, however it is computationally expensive. One of the shortcomings of the power method is that its convergence depends on magnitude of the highest eigenvalue and the next largest eigenvalue. The convergence will become delayed or possibly diverge if the ratio is tiny. When its dimensionality is greater than one, it produces only one pseudo-eigenvalue. We propose using inverse power iteration in this case, which can produce good results for high dimensional data. However, methods based on inverse power iteration can be quite expensive computationally, and the linear systems can be ill-conditioned when the data has outliers. We propose an improved algorithm, Inverse Power Iteration Clustering with preconditioning, that addresses the solutions of some of these issues. Moreover, Spectral Clustering can be used to implement pooling operations that combine nodes from the same cluster in Graph Neural Networks (GNNs). Data is clustered by enclosing data points in a low-dimensional subspace formed from the similarity matrix. We propose a pooling method as well as well-posed Laplacian with preconditioner for graph neural networks that address significant shortcomings of existing pooling operators and provide a deep learning methodology to overcome these constraints. Our proposed method combines the advantages of graph-theoretical methods with the adaptability of learnable techniques. We analyze and evaluate our approaches for unsupervised node clustering and supervised classification on a number of well-known benchmark datasets.