DOCTORAL THESIS

Numerical algorithms for data clustering

Liu, Ye

Date of Award:
2019

Abstract

Data clustering is a process of grouping unlabeled objects based on the imformation describing their relationship. And it has obtained a lot of attentions in data mining for its wide applications in life. For example, in marketing, companys are interested in finding groups of customers with similar purchase behavior, which will help them to make suitable plans to gain more profits. Besides, in biology, we can make use of data clustering to distinguish planets and animals given their features. Whats more, in earthquake analysis, by clustering observed earthquake epicenters, dangerous area can be identified, it would be helpful for people to take measures to protect them from earthquake in advance. In general, there isnt one clustering algorithm which can solve all the problems. Algorithms are specifically designed to analyze different data categories. In this thesis, we study several novel numerical algorithms for data clustering mainly applied on multi-view data and tensor data.

More accurate clustering result can be achieved on multi-view data by integrating information from multiple graphs. However, Most existing multi-view clustering method assume the degree of association among all the graphs are the same. One significant truth is some graphs may be strongly or weakly associated with other graphs in reality. Determining the degree of association between graphs is a key issue when clustering multi-view data. In Chapter 2, 3 and 4, we propose three different models to solve this problem.

In chapter 2, a block signed matrix is constructed to integrate information in each graph with association among graphs together. Then we apply spectral clustering on it to seek different cluster structure for each graph respectively and determine the degree of association among graphs using their own cluster structure at the same time. Numerical experiments including simulations, neuron activity data and gene expression data are conducted to illustrate the state-of-art performance of algorithm in clustering and graph association.

In Chapter 3, we further consider multiple graphs clustering with graph association solved by self-consistent field iterative algorithm. By using the block graph clustering framework, graphs association are considered to enhance clustering result, and then better clustering result would be used to calculate more accurate association. Self-consistent field iterative method is employed to solve this problem, and the convergence analysis

is also presented. Simulations are also carried out to demonstrate the outperformance of our method. Two gene expression data are used to evaluate the effectiveness of proposed model.

In Chapter 4, we formulate the multiple graphs clustering problem with the graph association as an objective function, and the graph association is considered as a term in the objective function. The proposed model can be solved efficiently by using gradient flow method. We also present its convergence analysis. Experiments on synthetic data sets and two gene expression data are given to show the efficiency in clustering and capability in graphs association.

In the last three chapters, we use multiple graphs to represent the multi-view data. A key challenge is high dimensionality when the number of graphs or objects is large-scale. Moreover, tensor is another common technique to describe multi-view data. Thus tensor decomposition method can be used to learn a low-dimensional representation for high dimensional data firstly and then perform clustering efficiently, which has attract worldwide attention of researchers. In Chapter 5, we propose an orthogonal nonnegative Tucker decomposition method to decompose high-dimensional nonnegative tensor into tensor with smaller size for dimension reduction, and then perform clustering analysis. A convex relaxation algorithm of the augmented Lagrangian function is devoloped to solve the optimization problem and the convergence of the algorithm is discussed. We employ our proposed method on several real image data sets from different real world application, including face recognition, image representation and hyperspectral unmixing problem to illustrate the effectiveness of proposed algorithm.

Keywords: Spectral clustering, Multi-view data, Multiple graphs, Multiple graphs clustering, Multiple graphs association, Signed graph, Unsigned graph, Nonlinear eigenvalue problem, Self-consistent field iteration, Gradient flow, Multiple orthogonal constraints, Iterative method, Nonnegative tensor, Tucker Decomposition, Orthogonality, Dimension reduction.

Table of Contents