

DOCTORAL THESIS

Numerical algorithms for data clustering

Liu, Ye

Date of Award:
2019

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

Data clustering is a process of grouping unlabeled objects based on the information describing their relationship. And it has obtained a lot of attentions in data mining for its wide applications in life. For example, in marketing, companies 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 developed 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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
Chapter 1 Introduction	1
1.1 Data Clustering	1
1.2 Graph based data clustering	2
1.2.1 Unsigned Graphs, Graph Laplacians and Graph Cut Theorem	3
1.2.2 Signed Graphs, Signed Laplacians and Signed Graph Cut Theorem	6
1.3 Multi-view Clustering	9
1.4 Nonnegative Tensor Decomposition for Clustering	11
1.5 Contribution and Outline	13
Chapter 2 Multiple Graphs Association and Clustering	16
2.1 Introduction	16
2.2 The Proposed Method	17
2.2.1 Block Laplacian Signed Matrices	17
2.2.2 Multi-structure Clustering by Eigenvectors	20
2.2.3 Consistency Weights Calculation	22

2.3	Experimental Results	24
2.3.1	Synthetic Data Sets	24
2.3.2	Real Data Sets	32
2.3.3	Neural Activity Data	32
2.3.4	Gene Expression Data: COAD-KIPAN-KIRC	36
2.3.5	Gene Expression Data: LUAD-LUSC-OV-UCEC	40
Chapter 3	Multiple Graphs Clustering by Self-Consistent Field Iterative Method	46
3.1	Introduction	46
3.2	The Proposed Model	47
3.2.1	Block Adjacency matrix with domain association	47
3.2.2	Multi-domain clustering by Self-consistent field iterative method .	49
3.2.3	Weight matrix calculation	49
3.3	Convergence Analysis	51
3.4	The general case	58
3.4.1	Convergence Analysis	60
3.5	Synthetic data	63
3.6	Gene expression Data	71
3.6.1	Gene expression data: LUAD-LUSC-OV-UCEC	71
3.6.2	Gene expression data: COAD-KIPAN-KIRC	73
Chapter 4	Multiple Graph Clustering by Gradient Flow Method	77
4.1	Introduction	77
4.2	The proposed algorithm	79
4.2.1	Gradient flow method	79
4.2.2	The discrete approximation algorithm and convergence analysis .	81
4.3	Numerical experiment	82
4.3.1	Synthetic data	82
4.3.2	Cancer gene expression data: LUAD-LUSC-UCEC	85
4.3.3	Cancer gene expression data: BRCA-OV-KIPAN-KIRC	90
Chapter 5	Image Clustering By Orthogonal Nonnegative Tucker Decomposition	94
5.1	Introduction	94

5.2	The Optimization Method	96
5.2.1	The Factor Matrix	97
5.2.2	The Algorithm and Convergence Analysis	100
5.2.3	The Core Tensor \mathcal{S}	101
5.3	Experimental Results	102
5.3.1	Feature Extraction and Face Recognition	102
5.3.2	Image Representation	104
5.3.3	Hyperspectral Unmixing	108
Chapter 6	Conclusion and Future Work	112
6.1	Conclusion	112
6.2	Future Work	113
	Curriculum Vitae	127