Some Algorithmic Studies in High-dimensional Categorical Data Clustering and Selection Number of Clusters

LI Junjie

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Principal Supervisor: NG. Michael Kwok-Po

Hong Kong Baptist University

October 2008
Abstract

Every day, people encounter a large amount of information and store or represent it as data, for further analysis and management. Data mining is a desirable technique to handle a huge amount of data. Data clustering is an important and applications-oriented branch of data mining. Its goal is to estimate the structure or density of a data set without a training signal. There are many data clustering methods that are very in their complexity and effectiveness, due to a wide number of applications that these algorithms can manage.

The aim of this thesis includes two parts. Firstly, we want to discuss some novel clustering algorithms for categorical data. K-modes algorithm used a simple frequency-based dissimilarity measure that there still has a room to improve. On the other hand, the techniques to deal with large volume, high-dimensionality categorical data are still virgin soils in text clustering. These techniques can be widely used in text, biomedical and transaction data. Secondly, the problem of determination of the number of clusters has remained open for decades since there is no appropriate theory for solving it except for some heuristic techniques. We are going to propose an innovative approach, based on a k-means type algorithm, to select the corrected number of clusters.

The major contributions of this thesis are as follows:

- A novel dissimilarity measure is proposed for k-modes Clustering Algorithm. The classical k-mode algorithm used a simple matching dissimilarity measure for categorical objects, modes instead of means for clusters, and a frequency-based method to update modes in the clustering process to minimize the clustering cost function. The distance between two objects computed with the simple matching similarity measure is either 0 or 1. Their main idea of the new dissimilarity measure is to use the relative attribute frequencies of the cluster modes in the similarity measure in the k-modes objective function. We give a rigorous proof, in Theorems 3 and 4, that the object cluster membership assignment method and the mode updating formulae under the new dissimilarity measure indeed minimize the objective function. With the formal proofs, we
assure that the modified $k$-modes algorithm can be used safely. Experiments have shown that the new dissimilarity measure is an important and effective method to improve the performance of classical $k$-mode algorithm.

- We originally design a novel subspace clustering algorithm for high-dimensional categorical data, for example the single nucleotide polymorphism (SNP) data sets. This algorithm is a $k$-modes type clustering method. In this algorithm, a new step is added in the $k$-modes clustering process to automatically calculate the weights of categorical dimensions in each cluster so that the important categorical dimension of each cluster can be identified by the weight values. The output of the algorithm includes two parts. It provides a partition of the data, so that the objects in each set of the partition constitute a cluster. In addition, it provides information to what dimensions are relevant for each partition. It’s shown that the algorithm converges to a local minimum of the objective function. The gain in performance is experimentally demonstrated.

- The last contribution of this thesis is to use the proposed *Agglomerative Fuzzy K-means Clustering Algorithm* to deal with the classical open problems in $k$-means type algorithm. The first problem is that the number of clusters $k$ needs to be determined in advance as an input to these algorithm, but it is unknown in the real data set. The second issue is the centers initialization problem. We propose an agglomerative fuzzy $k$-means clustering algorithm for numerical data to tackle the above two problems in application of the $k$-means-type clustering algorithms. The new algorithm can produce more consistent clustering results from different sets of initial clusters centers. Combined with cluster validation techniques, the new algorithm can determine the number of clusters in a data set. Experimental results have demonstrated the effectiveness of the new algorithm in producing consistent clustering results and determining the correct number of clusters in different data sets, some with overlapping inherent clusters.
# Contents

Declaration i  
Abstract ii  
Acknowledgements iv  
Table of Contents v  
List of Figures viii  
List of Tables xi  

**Chapter 1 Introduction**  
1.1 Clustering Techniques in Data Mining 3  
1.2 Motivations and Contributions 7  
1.3 Terms and Definitions 9  
1.4 Thesis Organization 9  

**Chapter 2 New Dissimilarity Measure in \( k \)-modes Clustering Algorithm**  
2.1 Introduction 12  
2.2 The \( k \)-modes Algorithm 13  
2.3 The New Dissimilarity Measure 16  
2.4 Experimental Results 20  
2.5 Conclusion 23
# Chapter 3 Subspace Clustering of High-dimensional Categorical Data

3.1 Introduction .................................. 25  
3.2 Related Work .............................. 27  
   3.2.1 Subspace Clustering Techniques ............... 27  
   3.2.2 Biclustering Algorithms ......................... 29  
3.3 Subspace Categorical Data Clustering ............. 30  
   3.3.1 The Algorithm .................................. 31  
   3.3.2 Convergency and Complexity Analysis .......... 34  
3.4 Experimental Results .......................... 36  
   3.4.1 Synthetic Data Sets ......................... 36  
   3.4.2 Results .................................. 38  
   3.4.3 Application to SNP Markers Detection .......... 43  
   3.4.4 Results on SNP Data Sets ..................... 46  
3.5 Conclusion .................................. 48

# Chapter 4 Agglomerative Fuzzy $K$-means Clustering Algorithm with Selection of Number of Clusters

4.1 Introduction .................................. 52  
4.2 Related Work .............................. 54  
   4.2.1 Cluster Validation .............................. 54  
   4.2.2 Optimization Functions ......................... 54  
   4.2.3 Competitive Agglomeration ..................... 55  
4.3 The Agglomerative Fuzzy $K$-means Algorithm ........ 56  
   4.3.1 The Optimization Procedure ..................... 57  
   4.3.2 The Properties of the Algorithm ............... 58  
   4.3.3 An Example .................................. 59  
   4.3.4 The Overall Implementation ..................... 62  
4.4 Experimental Results .......................... 63  
   4.4.1 Experiment 1 .................................. 65  
   4.4.2 Experiment 2 .................................. 69  
   4.4.3 Experiment 3 .................................. 72  
   4.4.4 Experiment 4 .................................. 75
Chapter 5 Conclusion  

Curriculum Vitae