DOCTORAL THESIS

Clustering of categorical and numerical data without knowing cluster number

Jia, Hong

Date of Award:
2013

Clustering of Categorical and Numerical Data without Knowing Cluster Number

JIA Hong

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

Principal Supervisor: Prof. Yiu-ming CHEUNG

Hong Kong Baptist University

April 2013

Abstract

Clustering is an effective technique for multivariate data analysis and is prevalent in different research areas. However, there are two challenging problems encountered in unsupervised clustering analysis. The first one is that many clustering algorithms need the number of clusters to be pre-assigned exactly; otherwise, they will almost always give out an incorrect clustering result. However, this vital information is not always available in practice. Besides, since most of the existing clustering approaches are applicable to purely numerical or categorical data only, but not the both, it becomes a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes as there exists an awkward gap between the similarity metrics for categorical and numerical data.

To handle the cluster number selection problem, in this thesis, we further study the penalization and cooperation mechanisms in competitive learning paradigm and propose a novel learning algorithm called Cooperative and Penalized Competitive Learning (CPCL), which implements the cooperation and penalization mechanisms simultaneously in a single competitive learning process. The integration of these two different kinds of competition mechanisms enables the CPCL to locate the cluster centers more quickly and be insensitive to the number of seed points and their initial positions. The promising experimental results on synthetic and real data demonstrate the superiority of the proposed algorithm.

Next, on the model selection for density mixture learning, we introduce the cooperation mechanism into the Maximum Weighted Likelihood (MWL) learning framework with a novel weight design and present an algorithm named Cooperative

EM (CEM) for mixture model learning with automatic model selection. Moreover, in order to enhance the robustness of the CEM algorithm to the initial parameters, we integrate the cooperation and penalization mechanisms together and accordingly generate a Cooperative and Penalized EM (CPEM) algorithm, in which the winning component in the competition at each time step will not only cooperate with the most promising rivals but also penalize some other rivals with a dynamic strength. It is found that the CPEM is insensitive to the initial parameters and can give a better estimation of the mixture model parameters, as well as the number of components. Experiments show the efficacy of the proposed algorithms on synthetic and real data.

Additionally, to address the problem of clustering on data mixed with categorical and numerical attributes, we present a general clustering framework based on the concept of object-cluster similarity and give a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. Accordingly, an iterative clustering algorithm is developed, whose outstanding performance is experimentally demonstrated on different benchmark data sets. Moreover, to circumvent the difficult problem of cluster number selection, we further develop a penalized competitive learning algorithm within the proposed clustering framework. The embedded competition and penalization mechanisms enable this improved algorithm to determine the number of clusters automatically by gradually eliminating the redundant clusters. The experimental results show the efficacy of the proposed approach.

Table of Contents