

## DOCTORAL THESIS

### Clustering of categorical and numerical data without knowing cluster number

Jia, Hong

*Date of Award:*  
2013

[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

# 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

<b>Declaration</b> . . . . .	<b>i</b>
<b>Abstract</b> . . . . .	<b>ii</b>
<b>Acknowledgements.</b> . . . . .	<b>iv</b>
<b>Table of Contents</b> . . . . .	<b>v</b>
<b>List of Tables.</b> . . . . .	<b>viii</b>
<b>List of Figures</b> . . . . .	<b>ix</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 K-means and Number of Clusters . . . . .	2
1.2 Learning Density Mixture Models and Model Selection Problem . . . . .	4
1.3 Clustering on Data with Categorical and Numerical Attributes . . . . .	6
1.4 Main Contributions of this Thesis . . . . .	7
1.5 Organization of the Thesis . . . . .	8
<b>Chapter 2 Literature Review of Related Works</b> . . . . .	<b>10</b>
2.1 Clustering without Knowing Cluster Number . . . . .	10
2.2 Model Selection for Density Mixture Learning . . . . .	14
2.3 Clustering on Data with Categorical and Numerical Attributes . . . . .	18

<b>Chapter 3</b>	<b>Cooperative and Penalized Competitive Learning for Robust Data Clustering . . . . .</b>	<b>21</b>
3.1	Introduction . . . . .	21
3.2	Overview of Existing Competitive Learning Models . . . . .	22
3.2.1	Overview of RPCL Algorithm and Its Variants . . . . .	22
3.2.2	Overview of CCL Algorithm . . . . .	25
3.3	Cooperative and Penalized Competitive Learning (CPCL) Approach .	26
3.3.1	Cooperation and Penalization Mechanisms in CPCL . . . . .	26
3.3.2	The CPCL Algorithm . . . . .	30
3.4	Comparisons between CPCL and Existing Counterparts . . . . .	31
3.5	Experimental Results . . . . .	34
3.5.1	Results on Synthetic Data . . . . .	35
3.5.2	Results on Real Data . . . . .	43
3.6	Summary . . . . .	49
<b>Chapter 4</b>	<b>Cooperative and Penalized EM Algorithm for Mixture Model Learning with Automatic Model Selection . . . . .</b>	<b>52</b>
4.1	Introduction . . . . .	52
4.2	Overview of MWL Learning Framework . . . . .	54
4.3	Cooperative EM algorithm . . . . .	56
4.3.1	Cooperative Mechanism . . . . .	56
4.3.2	The Proposed CEM Algorithm . . . . .	58
4.3.3	Experimental Results . . . . .	61
4.4	Cooperative and Penalized EM Algorithm . . . . .	63
4.4.1	Cooperative and Penalized Mechanism . . . . .	64
4.4.2	The Proposed CPEM Algorithm . . . . .	67
4.4.3	Experimental Results . . . . .	67
4.5	Comparative Study . . . . .	71
4.5.1	Experiment on Synthetic Data . . . . .	72
4.5.2	Real Data Set Analysis . . . . .	75

4.5.3	Color Image Segmentation . . . . .	77
4.6	Summary . . . . .	78
<b>Chapter 5</b>	<b>Categorical-and-Numerical-Attribute Data Clustering Based on a Unified Similarity Metric . . . . .</b>	<b>80</b>
5.1	Introduction . . . . .	80
5.2	Overview of K-prototype and K-modes Algorithms . . . . .	82
5.3	Clustering Problem and Object-cluster Similarity Metric . . . . .	85
5.3.1	Similarity Metric for Mixed Data . . . . .	85
5.3.2	Object-cluster Similarity Metric . . . . .	90
5.4	Iterative Clustering Algorithm . . . . .	91
5.5	Automatic Selection of Cluster Number . . . . .	94
5.5.1	Competition Mechanism . . . . .	95
5.5.2	Penalization Mechanism . . . . .	96
5.6	Experiments . . . . .	100
5.6.1	Performance Evaluation of OCIL Algorithm . . . . .	101
5.6.2	Performance Evaluation of PCL-OC Algorithm . . . . .	108
5.7	Summary . . . . .	110
<b>Chapter 6</b>	<b>Conclusions and Future Work . . . . .</b>	<b>112</b>
6.1	Conclusions . . . . .	112
6.2	Future Work . . . . .	114
	<b>Bibliography . . . . .</b>	<b>116</b>
	<b>Curriculum Vitae . . . . .</b>	<b>132</b>