

Faculty of Engineering and Information Technology
University of Technology, Sydney

Coupling Analysis in Educational Data

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

by

Mu Li

December 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Longbing Cao for his continuous support, patience, motivation, enthusiasm, and immense knowledge. His guidance was tremendously valuable over the course of my PhD research. I could not have had a better advisor and mentor.

My sincere thanks is also extended to Dr. Jinjiu Li for providing me with his continuous support over the duration of my PhD. Without his professional guidance and consistent help, this thesis would not have been possible.

Many thanks to my fellow labmates at the Advanced Analytics Institute: Chuming Liu, Yin Song, Junfu Yin for the stimulating discussions, and sleepless nights that we shared whilst working together on a range of deadlines, and for all the fun we had over the last four years.

My sincere appreciation goes to the team leader, Zhigang Zheng for his kindness and expert help with the project. I am also grateful to my team members, Allen Lin, Chunming Liu and Wei Cao, for their hard work in the project teams.

Finally, my deep and abiding gratitude is extended to my wife and my parents, for their unconditional love and support over the course of my PhD candidature.

Mu Li

March 2015 @ UTS

Contents

Certificate	i
Acknowledgment	iii
List of Figures	ix
List of Tables	xi
Abstract	xiii
Chapter 1 Introduction	1
1.1 Background	1
1.1.1 Coupling Analysis	2
1.1.2 Educational Data Analysis	3
1.1.3 Coupling Analysis in Educational Data	7
1.2 Research Issues and Objectives	9
1.3 Research Contributions	11
1.3.1 Coupled Similarity Based K-centroid Classifier	12
1.3.2 Coupled Similarity Based Pairwise SVM Classifier	12
1.3.3 Coupled Term to Term Relations Extractor	13
1.3.4 Large Scale Coupling Analysis Framework	14
1.4 Thesis Organization	15
Chapter 2 Literature Review and Foundation	18
2.1 Foundation	18
2.1.1 Coupled Similarity for Categorical Data	18
2.1.2 Classification Method	21
2.1.3 Text Mining	30

2.2	Literature Review of Education Data Mining	34
2.2.1	Student Performance Prediction	37
2.2.2	Sentiment Analysis	40
2.2.3	Segmentation of Different Users	42
2.2.4	Diversity of Scenarios	44
2.2.5	Technologies and Applications	56
2.3	Summary	66
Chapter 3 A Coupled Similarity based K-Nearest Centroid		
Classifier for Student Performance Prediction . . 68		
3.1	Introduction	68
3.2	Problem Statement	71
3.3	Coupled Similarities Based Classification	72
3.3.1	Clustering Within the Class with Coupled Similarities .	73
3.3.2	Spherical K-Means Clustering with Coupled Similarities	74
3.3.3	Classification with Coupled Similarities Weighted Cluster Centroid	75
3.4	Experiment and Evaluation	77
3.5	Application	81
3.5.1	Data Source and Pre-process	82
3.5.2	Classification Efficiency Comparison	85
3.6	Summary	87
Chapter 4 A Coupled Similarity Kernel based Pairwise Support Vector Machines for Student Performance Prediction 88		
4.1	Introduction	88
4.2	Problem Statement	91
4.3	Pairwise SVM with Coupled Similarity	93
4.3.1	Coupled Similarity as a Kernel	93
4.3.2	Pairwise SVM	94
4.4	Experiment and Evaluation	97

4.5	Application	100
4.6	Summary	105
Chapter 5 Extracting Coupling Relations from Term to Term for Student Social Media Sentiment Analysis . 106		
5.1	Introduction	106
5.2	Vector Presentation of Terms	111
5.2.1	Intra-Coupling Relation	111
5.2.2	Inter-Coupling Relation	114
5.3	Coupled Similarity based Document Classification	116
5.4	Experiment and Evaluation	117
5.4.1	Public Data Sets	117
5.4.2	Experiment	119
5.5	Application	122
5.6	Summary	123
Chapter 6 Large-Scale Coupling Analysis For Educational Data 125		
6.1	Introduction	125
6.2	Problem Statement	129
6.3	Parallelization	132
6.3.1	Spark Parallel Platform	132
6.3.2	Parallel Coupled Similarity	134
6.4	Integration	137
6.4.1	Parallel Coupled Similarity Based K-Means Algorithm	137
6.4.2	Parallel Coupled Similarity Based KNN Algorithm	138
6.5	Experiments and Evaluation	139
6.5.1	Experiment Setting	139
6.5.2	The Scalability of the Spark Coupled Similarity	141
6.5.3	The Performance of SK-Means	144
6.5.4	The Performance of SKNN	147
6.6	Application	147

CONTENTS

6.7 Summary	151
Chapter 7 Conclusions and Future Work	152
7.1 Conclusions	152
7.2 Future Work	153
7.2.1 Algorithms	153
7.2.2 Applications	154
Appendix	155
7.1 List of Publications	156
Bibliography	159

List of Figures

1.1	The Work Flow of Educational Data Analysis	4
1.2	Application of Student Performance Prediction	6
1.3	Application of Student Sentiment Analysis	7
1.4	The Profile of the Research Work of This Thesis	17
2.1	Publication Trend of the Educational Data Mining	57
3.1	Comparison Times with and without Clustering	73
3.2	The Training Data Set After Clustering	76
3.3	Classification Precision Comparison	79
3.4	Classification Recall Comparison	80
3.5	Classification F1 Measure Comparison	81
3.6	Data Source for Student Performance Prediction	84
3.7	Precision on Different Number of Clusters	86
3.8	Time Cost on Different Number of Clusters	86
4.1	Classification Precision Comparison	98
4.2	Classification Recall Comparison	99
4.3	Classification F1 measure Comparison	99
4.4	Subject Introduction to Electrical Engineering Classification Comparison	100
4.5	Subject Electronics and Circuits Classification Comparison . .	101
4.6	Subject Fundamentals of Electrical Engineering Classification Comparison	102

LIST OF FIGURES

4.7	Subject Circuit Analysis Classification Comparison	103
4.8	Subject Signals and Systems Classification Comparison	104
4.9	Two Strategies on All Subjects Classification Comparison	104
5.1	Term Frequency in Reuters-21578 R8	108
5.2	Sliding Window	111
5.3	Inter Relation from Term to Term	115
5.4	Classification (20 Newsgroup)	118
5.5	Classification (Reuters-21578 R52)	120
5.6	Classification (Reuters-21578 R8)	121
5.7	Different Window Type	122
5.8	The Data Flow of Student Sentiment Analysis	123
5.9	Classification Performance of Student Sentiment Analysis	124
6.1	The Overview of the Spark Cluster.	134
6.2	The Work Flow of Parallel Intra Coupled Similarity	135
6.3	The Work Flow of Parallel Pre-Inter Coupled Similarity	136
6.4	The Work Flow of Parallel Inter Coupled Similarity	137
6.5	Coupled Similarity Based Mean Step for Clustering	139
6.6	The Running time with different data sets.	142
6.7	The Running time with different feature numbers.	143
6.8	The Running time with different cores on 3 data sets.	144
6.9	The Purity Results on 10^4 instances.	145
6.10	The Purity Results on 10^5 instances.	146
6.11	The Purity Results on 10^6 instances.	146
6.12	The Accuracy Results on 10^5 instances.	148
6.13	The Accuracy Results on 10^5 instances.	148
6.14	The Accuracy Results on 10^6 instances.	149
6.15	Time Cost via Different Number of CPU Cores(Student Performance)	150
6.16	Time Cost via Different Number of CPU Cores(Student Sentiment)	150

List of Tables

1.1	Student Information Table	8
3.1	Coupled Similarity Between Objects	74
3.2	Comparison on Student Data	85
5.1	Most Related Terms of the Key-Term	113
6.1	Experimental Data Sets	140
6.2	Efficiency on Full Data Sets	141

Abstract

Educational data analysis refers to techniques, tools, and research designed to automatically extract meaning from large repositories of data generated by or related to people's learning activities in educational environments. It is a research field which focus on helping policymakers and administrators understand how analytics and data mining can be applied for the purposes of educational improvement. Unfortunately, most research on educational data only by applying the existing machine learning or data mining algorithms, very few publications have discussed the character of the data itself. Traditional data mining algorithms have disadvantages, in that most of them assume the independent and identically distributed (IID) of data objects, attributes, and values. However, real world data usually contains strong couplings among values, attributes and data objects, and this represents a considerable challenge to existing methods and tools. This thesis focuses on utilizing coupling analysis in educational data analysis tasks. In particular, it focuses on two educational data analysis tasks: student performance prediction, and student social media sentiment analysis.

The student performance prediction task is firstly examined. This thesis begins with the most straightforward method which integrates coupling similarities as the distance for a weighted k-nearest centroid classifier. This method considers not only the intra-coupled similarity within an attribute but also the inter-coupled similarity between attributes. Computational cost is high for coupling analysis. Hence, a more efficient method is proposed that selects the centroid objects instead of all objects in the nearest neigh-

bor search process. Furthermore, integrating support vector machines with coupled similarity. The original SVMs is designed for numerical data. This thesis develops a novel pairwise SVMs that use the coupled similarity metric as a kernel between data objects with nominal attributes. The experimental result shows the two proposed methods outperform the traditional SVMs and other popular classification methods on various public data sets, and the student performance prediction task.

Secondly, the student social media sentiment analysis is examined. Unlike linguistic methods, this thesis learns how to classify student sentiment by applying data mining on the labeled historical data. Most previous research employs the vector-space model for text representation and analysis, however, the vector-space model does not utilize the information about the term to term relationships. In other words, the traditional text mining techniques assume the relations between term to term are independent and identically distributed (IID). This thesis introduces a novel term representation by involving coupling relations between neighbors. This coupling representation provide much richer information which enables us to create a coupled similarity metric from document to document, and a coupling document similarity based k-nearest centroid classifier applied to the classification task. Experiments verify that the proposed approach outperforms the classic vector-space based classifier and displays distinct advantages and richness in terms of student social media sentiment analysis tasks.

Finally, due to the complexity of the proposed algorithm and the enormous amount of the educational related data source, a scalable educational data mining platform is in great demand. Hence, with the help of the Spark cluster, a novel coupling similarity based learning approach has been proposed to cater for the big data learning problem by parallelizing the coupled similarity calculation process. Further, the parallel k-NN for classification and k-Means for the clustering task has been proposed. Compared to the original algorithms, the experimental results show that the proposed methods not only outperforms the clustering and classification performance of the

baselines, but also represent a huge improvement on the data scale in terms of the time efficiency. Accordingly, the proposed framework has already been implemented, a scalable educational data analysis platform with coupling analysis will serve to meet a host of future challenges.

