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 consistant help, this thesis would not have been possible.

Many thanks to my fellow labmates at the Advanced Analystics 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.

Fianlly, 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

List of Figures

List of Tables

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 experiment 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.