Faculty of Engineering and Information Technology
University of Technology, Sydney

Debt Detection and Debt Recovery with Advanced Classification Techniques

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

by

Shanshan Wu

October 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

I would like to express my sincere gratitude to my supervisor, Dr. Ling Chen, for the continuous support of my Ph.D thesis, whose expertise, understanding, and patience added considerably to my thesis completion. I attribute the level of my degree to her encouragement and effort and without her this thesis, too, would not have been completed or written.

My sincere thanks also goes to my co-supervisor, Prof. Chengqi Zhang, for offering me the scholarship opportunities and leading me working on the business project in Centrelink. His advice on both research as well as on my career have been priceless. He has supported me thoughout my Ph.D study with his patience and knowledge whilst allowing me the room to work in a flexible time frame.

This thesis was supported by the Australian Research Council (ARC) Linkage Project between University of Technology, Sydney, Australia and Centrelink Australia. I worked on site in Centrelink Canberra for about one and half years. It is one of the most enjoyable working experience I've ever had. A very special thanks goes out to Dr. Yanchang Zhao and Dr. Huaifeng Zhang, who provided me with direction, shared their truthful and illuminating views on a number of issues and became more of a mentor and friend. I must also acknowledge Mr. Hans Bohlscheid, Mr. Peter Newbigin and Mr. Brett Clark from Business Integrity Programs Branch, Centrelink, Australia for their support of domain knowledge and helpful suggestions.

Last but not the least, I would like to thank my family for the support they provided me through my entire life. In particular, I must acknowledge

my mum and my husband, without whose love, encouragement and support, I would not have finished this thesis. At the end I would like express appreciation to my beloved son, Jayden, whose birth lights up my life and makes me more powerful to complete the thesis.

Contents

List of Figures

List of Tables

Abstract

My study is part of an ARC linkage project between University of Technology, Sydney and Centrelink Australia, which aims to applying data mining techniques to optimise the debt detection and debt recovery. A debt indicates an overpayment made by the government to a customer who is not entitled to that payment.

In social security, an interaction between a customer and the government department is recorded as an activity. Each customer's activities happen sequentially along the time, which can be regarded as a sequence. Based on the experience of debt detection experts, there are usually some patterns in the sequence of activities of customers who commit debts. The patterns indicating the customers' intention to be overpaid can thus be used to discover or predict debt occurrence. The development of debt detection and recovery over sequential transaction data, however, is a challenging problem due to following reasons. (1) The size of transaction data is vast, and the transaction data are being generated continuously as the business goes on. (2) Transaction data are always time stamped by the business system, and the temporal order of the transaction data is highly related to the business logic. (3) The patterns and relationships hidden behind the transaction data may be affected by a lot of factors. They are not only dependent on business domain knowledge, but also subject to seasonal and social factors outside the business. Based on a survey of existing methods on debt detection and recovery, data mining techniques are studied in this thesis to detect and recovery debt in an adaptive and efficient fashion.

Firstly, sequence data is used to model the evolvement of customer activities, and the sequential patterns generalize the trends of sequences. For long running sequence classification issues, even if the sequences come from the same source, the sequential patterns may vary from time to time. An adaptive sequential classification model is to be built to make the sequence classification adapt to the sequential pattern variation. The model is applied to 15,931 activity sequences from Centrelink which includes 849,831 activity records. The experimental results show that the proposed adaptive sequence classification framework performs effectively on the continuously arriving data.

Secondly, a new technique of sequence classification using both positive and negative patterns is to be studied, which is able to find the relationship between activity sequences and debt occurrences and also the impact of oncoming activities on the debt occurrence. The same dataset is used for the evaluation. The outcome shows if built with the same number of rules, in terms of recall, the classifier built with both positive and negative rules outperforms traditional classifiers with only positive rules under most conditions.

Finally, decision trees are to be built in the thesis to model debt recovery and predict the response of customers if contacted by phone. The customer contact strategy driven by the model aims to improve the efficiency of debt recovery process. The model is utilized in a real life pilot project for debt recovery in Centrelink. The pilot result outperforms the traditional random customer selection.

In summary, this thesis studies debt detection and debt recovery in social security using data mining techniques. The proposed models are novel and effective, showing potentials in real business.