Faculty of Engineering and Information Technology

University of Technology, Sydney

SOCIAL SECURITY DATA MINING: AN AUSTRALIAN CASE STUDY

A Thesis Submitted in Fulfilment of the Requirements for The Degree of Doctor of Philosophy, Faculty of Engineering and Information Technology, University of Technology Sydney.

By

Hans Michael Bohlscheid

October 2013

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I, Hans Michael Bohlscheid, 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

_____

Hans M Bohlscheid

01 /10 /2013

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to Professor Longbing Cao for his supervision and ongoing support over many years. I regard Professor Cao as an energetic, inspiring individual and will always remember our many discussions and his spontaneous ideas.

I also remain grateful to my colleagues and friends Drs Huaifeng Zhang and Yanchang Zhao for their advice regarding the technical aspects of my thesis.

In the mix of the many people to whom I am indebted are Yuri Zubrutsky, Daniel Marlay and Richard Brookes who independently evaluated my results and last but not least, I extend my heartfelt thanks to my dear colleagues Peter Newbigin and Brett Clark for their expertise in data extraction and their acute knowledge of Centrelink business practices.

In their own unique way, each of the above contributed to the enjoyment and success of my research and collectively, they are responsible for the skills and subject matter knowledge I possess today.

"I was striking an uneasy balance between the ambition I had for myself, and what those closest to me expected of me. So I stopped pretending to myself that I was anything other than what I was, and began to direct all my energy into finishing the only work that mattered to me."

J.K. Rowling

TABLE OF CONTENTS

SOCIAL SECURITY DATA MINING: AN AUSTRALIAN CASE STUDY 1

CERTIFICATE OF AUTHORSHIP/ORIGINALITY 2

ACKNOWLEDGEMENT 3

TABLE OF CONTENTS 4

LIST OF FIGURES 9

LIST OF TABLES 10

ABSTRACT 12

1 INTRODUCTION 14

1.1 BACKGROUND 14

1.2 BUSINESS NEEDS AND RESEARCH ISSUES 15

1.2.1 Business Needs ：Social Security Service Delivery, Reform and Challenge 15

1.2.2 Research Issues in Social Security Data Mining 17

1.3 RESEARCH METHODOLOGY 19

1.3.1 Positive / Negative Sequential Rules Mining for Debt-oriented Analysis 19

1.3.2 Sequence Classification to Predict Debt-related Activities 19

1.3.3 Combined Mining for Social Security Data 20

1.4 THESIS CONTRIBUTION 20

1.5 THESIS ORGANIZATION 21

2 REVIEW ON SOCIAL SECURITY/WELFARE DATA MINING RESEARCH 22

2.1 INTRODUCTION 22

2.2 SOCIAL SECURITY SERVICES AND DATA 22

2.2.1 Social Security Business 22

2.2.2 Social Security Data 24

2.3 COMPREHENSIVE PICTURE 27

2.4 TECHNICAL PERSPECTIVE 28

2.5 RELATED WORK OF SOCIAL SECURITY DATA MINING 30

2.6 RETROSPECTION ON MINING SOCIAL SECURITY DATA 30

3 SOCIAL SECURITY DATA MINING FRAMEWORK 33

3.1 A BASIC FRAMEWORK 33

3.2 SOCIAL SECURITY DATA MINING GOALS 35

3.3 SOCIAL SECURITY DATA MINING CHALLENGES 37

3.3.1 Social Security Data Processing 37

3.3.2 Social Security Pattern Analysis 39

3.3.3 Knowledge Delivery 40

3.4 SOCIAL SECURITY DATA MINING TASKS 42

3.4.1 Data Processing 42

3.4.2 Activity Analysis 43

3.4.3 Customer Risk Analysis 44

3.4.4 Earnings Analysis 45

3.4.5 Change Detection 45

3.4.6 Payment Accuracy Analysis 46

3.4.7 Fraud Detection 48

4 POSITIVE/NEGATIVE SOCIAL SECURITY SEQUENTIAL RULES ANALYSIS 49

4.1 POSITIVE/NEGATIVE SEQUENTIAL RULES 49

4.1.1 Introduction *49*

4.1.2 Background and Related Work *50*

4.1.3 Problem Statement *52*

4.2 EFFICIENT MINING OF EVENT-ORIENTED NEGATIVE SEQUENTIAL RULES 54

4.2.1 Algorithm for Mining Event-Oriented Negative Sequential Rules *54*

4.2.2 Experimental Evaluation *56*

4.2.3 Conclusion *59*

4.3 MINING BOTH POSITIVE AND NEGATIVE IMPACT-ORIENTED SEQUENTIAL RULES FROM TRANSACTIONAL DATA 60

4.3.1 Mining Impact-Oriented Sequential Rules *60*

4.3.2 Experimental Results *62*

4.3.3 Conclusions *66*

5 PREDICT DEBT-RELATED SOCIAL SECURITY ACTIVITY SEQUENCES 67

5.1 PROBLEM STATEMENT OF SEQUENCE CLASSIFICATION 67

5.2 SEQUENCE CLASSIFICATION USING BOTH POSITIVE AND NEGATIVE SEQUENTIAL PATTERNS 68

5.2.1 Discriminative Sequential Patterns *68*

5.2.2 Building Sequence Classifiers *69*

5.2.3 Case Study *70*

5.3 DEBT DETECTION IN SOCIAL SECURITY BY ADAPTIVE SEQUENCE CLASSIFICATION 76

5.3.1 Discriminative Frequent Patterns Boosting *76*

5.3.2 Adaptive Sequence Classification Framework *78*

5.3.3 Case Study *80*

5.4 CUSTOMER ACTIVITY SEQUENCE CLASSIFICATION FOR DEBT PREVENTION IN SOCIAL SECURITY 85

5.4.1 Interestingness Measure 85

5.4.2 Sequence Classification 86

5.4.3 Case Study 90

6 MINING COMBINED SOCIAL SECURITY PATTERNS 94

6.1 INTRODUCTION 94

6.2 RELATED WORK 94

6.3 THE PROBLEM 95

6.4 COMBINED PATTERN MINING 98

6.4.1 Definitions of Combined Patterns 98

6.4.2 Interestingness Measures for Combined Patterns 99

6.4.3 Redundancy in Combined Patterns 102

6.5 A CASE STUDY 103

7 RARE CLASS ASSOCIATION RULE MINING WITH MULTIPLE IMBALANCED ATTRIBUTES 107

7.1 BACKGROUND 107

7.1.1 Class Association Rules 108

7.1.2 Data Imbalance in Association Rule Mining 109

7.2 NOVEL ASSOCIATION RULE MINING PROCEDURE 109

7.2.1 Interestingness Measures 110

7.2.2 Transformation 112

7.3 TEST CASE 113

7.3.1 Datasets Involved 113

7.3.2 Experimental Results *114*

7.4 CONCLUSIONS 117

8 CONCLUSIONS AND FUTURE WORK 118

8.1 CONCLUSIONS 118

8.2 FUTURE WORKS 118

8.2.1 Data Mining Applications in Social Security *118*

8.2.2 Sequential Rules Mining and Sequence Classification *119*

8.2.3 Development of Further Models *119*

8.2.4 A Straightforward Approach to Ongoing Research and Development *120*

LIST OF PUBLICATIONS 121

AWARDS 121

BOOK CHAPTERS 121

CONFERENCE PAPERS 121

JOURNAL ARTICLES 122

REFERENCES 123

LIST OF FIGURES

FIGURE 2-1 SOCIAL SECURITY BUSINESS WORKFLOW 23

FIGURE 3-1 SSDM FRAMEWORK 33

FIGURE 3-2 SSDM GOALS 35

FIGURE 3-3 SSDM CHALLENGES 37

FIGURE 3-4 SSDM TASKS 42

FIGURE 4-1 ALGORITHM FOR DISCOVERING EVENT-ORIENTED NEGATIVE SEQUENTIAL RULES 56

FIGURE 4-2 SCALABILITY WITH MINIMUM SUPPORT 57

FIGURE 4-3 SCALABILITY WITH THE NUMBER OF SEQUENCES 58

FIGURE 4-4 SCALABILITY WITH THE NUMBER OF ITEMS PER SEQUENCE 58

FIGURE 4-5 SCALABILITY WITH THE AVERAGE LENGTH OF PATTERNS 59

FIGURE 4-6 PSEUDOCODE FOR DISCOVERING IMPACT-ORIENTED NEGATIVE SEQUENTIAL RULES 61

FIGURE 4-7 SCALABILITY WITH (A) SUPPORT; (B) THE NUMBER OF SEQUENCES; AND (C) THE LENGTH OF SEQUENCES 63

FIGURE 4-8 A GROWING SEQUENTIAL PATTERN “ADV ADV CCO” 66

FIGURE 5-1 ARCHITECTURE OF ADAPTIVE SEQUENCE CLASSIFICATION 79

FIGURE 5-2 ADAPTIVE CLASSIFICATION MODEL 80

FIGURE 5-3 EFFECTIVENESS OF DISCRIMINATIVE PATTERNS BOOSTING 82

FIGURE 5-4 ROC CURVES OF ADAPTIVE SEQUENCE CLASSIFICATION FRAMEWORK 84

FIGURE 5-5 HIERARCHICAL SEQUENCE CLASSIFICATION ALGORITHM 88

FIGURE 7-1 PROPOSED ALGORITHM 110

FIGURE 7-2 DISTRIBUTION OF THE IMBALANCED ATTRIBUTES 115

LIST OF TABLES

TABLE 2-1 CENTRELINK BUSINESS DIMENSIONS 2008–2009 24

TABLE 4-1 NOTATIONS 50

TABLE 4-2 SUPPORTS, CONFIDENCES AND LIFTS OF FOUR TYPES OF SEQUENTIAL RULES 54

TABLE 4-3 SELECTED POSITIVE AND NEGATIVE SEQUENTIAL RULES 65

TABLE 5-1 FEATURE-CLASS CONTINGENCY TABLE 69

TABLE 5-2 EXAMPLES OF ACTIVITY TRANSACTION DATA 71

TABLE 5-3 SELECTED POSITIVE AND NEGATIVE SEQUENTIAL RULES 72

TABLE 5-4 THE NUMBER OF PATTERNS IN PS10 AND PS05 73

TABLE 5-5 CLASSIFICATION RESULTS WITH PATTERN SET PS05-4K 75

TABLE 5-6 CLASSIFICATION RESULTS WITH PATTERN SET PS05-8K 75

TABLE 5-7 CLASSIFICATION RESULTS WITH PATTERN SET PS10-4K 75

TABLE 5-8 CLASSIFICATION RESULTS WITH PATTERN SET PS10-8K 76

TABLE 5-9 THE NUMBER OF PATTERNS IN THE FOUR PATTERN SETS 76

TABLE 5-10 CENTRELINK DATA SAMPLE 81

TABLE 5-11 DATA WINDOWS 83

TABLE 5-12 2 BY 2 FEATURE-CLASS CONTINGENCY TABLE 86

TABLE 5-13 SAMPLES OF CENTRELINK ACTIVITY DATA 91

TABLE 5-14 PERFORMANCE OF DIFFERENT ALGORITHMS 92

TABLE 5-15 COMPARISON OF THE PROPOSED ALGORITHM TO CONVENTIONAL ALGORITHM 93

TABLE 6-1 TRANSACTIONAL DATA 96

TABLE 6-2 CUSTOMER DEMOGRAPHIC DATA 96

TABLE 6-3 TRADITIONAL ASSOCIATION RULES 96

TABLE 6-4 COMBINED ASSOCIATION RULES 96

TABLE 6-5 COMBINED RULE PAIRS 97

TABLE 6-6 TRADITIONAL ASSOCIATION RULES 104

TABLE 6-7 SELECTED COMBINED RULES 104

TABLE 6-8 SELECTED COMBINED RULE CLUSTERS 105

TABLE 7-1 SELECTED RESULTS WITH BALANCED ATTRIBUTES 114

TABLE 7-2 SELECTED RULES WITH IMBALANCED ATTRIBUTES CAPTION STYLE 116

TABLE 7-3 SELECTED RESULTS OF THE COMBINED ASSOCIATION RULES 116

Abstract

Data mining in business applications has become an increasingly recognized and accepted area of enterprise data mining in recent years. In general, while the general principle and methodologies of data mining and machine learning are applicable for any business applications, it is often essential to develop specific theories, tools and systems for mining data in a particular domain such as social security and social welfare business. This necessity has led to the concept of *social security and social welfare data mining*, the focus of this thesis work.

Social security and social welfare business involves almost every citizen's life at different life periods. It provides fundamental and crucial government services and support to varied populations of specific need. A typical scenario in Australia is that it not only connects one third of our populations, but also associates with many relevant stakeholders, including banking business, taxation and Medicare. Such business engages complicated infrastructure, networks, mechanisms, policies, activities, and transactions. Data mining of such business is a brand new application area in the data mining community.

Mining such social welfare business and data is challenging. The challenges come from the unavailable benchmark and experience in the data mining for this particular domain, the complexities of social welfare business and data, the exploration of possible doable tasks, and the implementation of data mining techniques in relation to the business objectives.

In this thesis, which adopts a practice-based innovative attitude and focusses on the marriage of social welfare business with data mining, we believe we have realised our objective of providing a systematic and comprehensive overview of the social security and social welfare data mining. The main contributions consist of the following aspects:

- As the first work of its kind, to the best of our knowledge, we present an overall picture of social security and social welfare data mining, as a new domain driven data mining application.
- We explore the business nature of social security and social welfare, and the characteristics of social security data.

- We propose a concept map of social security data mining, catering for main complexities of social welfare business and data, as well as providing opportunities for exploring new research issues in the community.
- Several case studies are discussed, which demonstrate the technical development of social security data mining, and the innovative applications of existing data mining techniques.

The nature of social welfare is spreading widely across the world in both developed and developing countries. This thesis work therefore is timely and could be of important business and government value for better understanding our people, our policies, our objectives, and for better services of those people of genuine needs.