

**FINANCIAL FRAUD DETECTION USING MACHINE
LEARNING**

Thilini Upeksha Kodituwakku

(148225U)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

**University of Moratuwa
Sri Lanka**

March 2018

FINANCIAL FRAUD DETECTION USING MACHINE LEARNING

Thilini Upeksha Kodituwakku

(148225U)

**Thesis submitted in partial fulfillment of the requirements for the
degree Master of Science in Computer Science**

Department of Computer Science and Engineering

**University of Moratuwa
Sri Lanka**

March 2018

Declaration

I hereby declare that this is my own work and this thesis/dissertation does not incorporate any material previously submitted for a Degree or Diploma in any other University or institute of higher learning without proper acknowledgement and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Masters/MPhil/PhD thesis/ Dissertation under my supervision.

Signature of the supervisor:

Date:

Prof. Amal Shehan Perera

ABSTRACT

The method of performing transactions by means of payment cards is extremely efficient and the payment card industry is rapidly growing in popularity. However, the frauds associated with the payment cards are increasing and the patterns are evolving. Although a relatively smaller percentage is detected, fraud has become a major issue that affects the global banking industry. Machine learning techniques are widely used for payment card fraud detection.

The use of machine learning techniques generates successful results as there are large numbers of historical data that could be used for mining and manipulation. There are various machine learning algorithms available to construct fraud detection models. The main drawback of those models is their inability to deliver results accurately and efficiently at the level required by the industry as there is only a fine line between the fraudulent and non-fraudulent transactions.

The aim of this research is to create a model that reduces the present gap in the detection of payment card frauds using the ensemble machine learning technique. Ensemble methods are learning models that achieve performance by combining the opinions of multiple weaker models. The performance evaluation of a new ensemble model has been done on the real world financial data and the results indicated its capability of identifying a high percentage of frauds with low false alarm rate than the existing models in the payment card industry. Finally, results are analyzed, interpreted and directions for further research are suggested.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof. Amal Shehan Perera for providing me with the knowledge, support, and guidance throughout the duration of this project. Furthermore, I would like to thank my parents for the continuous support they gave me during this period. Also, I like to thank my colleagues who helped me immensely to gain an in-depth knowledge of payment card industry.

Table of Contents

ABSTRACT.....	iii
Acknowledgement	iv
Table of Contents	v
List of Figures	ix
List of Tables	x
List of Abbreviations	xi
1. INTRODUCTION	1
1.1 Problem Statement	3
1.2 Motivation	3
1.3 Objective	4
2. LITERATURE REVIEW	5
2.1 Data mining	7
2.2 Supervised learning	8
2.2.1 Classification	8
2.2.1.1 K-nearest neighbour classifier	9
2.2.1.2 Case-based reasoning.....	9
2.2.1.3 Rule-based approach.....	10
2.2.1.4 Fuzzy set approach.....	11
2.2.1.5 Artificial Neural Network (ANN).....	11
2.2.1.6 Bayesian network.....	13
2.2.1.7 Decision trees	14
2.2.1.8 Genetic algorithm.....	14
2.3 Unsupervised learning.....	16
2.3.1 Clustering.....	16
2.3.1.1 Partitioning method.....	17

2.3.1.2 Hierarchical method.....	18
2.3.1.3 Density-based method.....	19
2.3.1.4 Conceptual method	19
2.4 Comparison of approaches	20
2.5 Ensemble methods.....	20
2.5.1 Bagging.....	21
2.5.2 Boosting	22
2.5.3 Arcing	23
2.5.4 Random trees	23
2.5.5 Random forests	24
2.6 Attribute selection	25
2.7 Challenges	26
2.7.1 Skewed distribution	26
2.7.2 Handling noise	26
2.7.3 Overlapping of data	27
2.7.4 Handling new types of frauds (Concept Drift)	28
2.7.5 Identification of good metrics.....	28
3. METHODOLOGY	29
3.1 Objective	29
3.2 Approach	29
3.3 Scope	30
3.4 Data selection	31
3.4.1 Dataset	32
3.5 Data pre-processing.....	32
3.6 Splitting the data.....	33

3.7	Training data	33
3.8	Levenberg – Marquardt algorithm	35
3.8.1	First-order optimization algorithms	36
3.8.2	Second-order optimization algorithms	36
3.9	Implementation and Methodology	39
3.9.1	User Entered Rules	39
3.9.2	Rule-based classification	40
3.9.3	RIPPER algorithm	41
3.9.4	Bagging with Levenberg-Marquardt neural networks algorithm	45
3.10	Implementation.....	46
4.	RESULTS AND EVALUATION.....	47
4.1	Goal	47
4.2	Testing Strategy.....	47
4.2.1	Specified Rules	48
4.2.1.1	Rules	49
4.3	Test results analysis.....	50
4.3.1	Evaluation of the results of three test datasets.....	50
4.3.2	Comparison of classification accuracy of the model with base classifiers	54
4.3.3	Review of accuracy of the model	56
4.3.4	Comparison of efficiency of the model with base classifiers	57
4.4	Performance comparison of the model with counterparts.....	59
4.4.1	Accuracy	60
4.4.1.1	Testing for significance of accuracy	61
4.4.2	Efficiency.....	63
5.	CONCLUSION AND FUTURE WORK	66

6. REFERENCES	67
7. APPENDICES	73
Appendix A – Steps of implementation of RIPPER algorithm	73
Appendix B – Request and response messages of FDM	74

List of Figures

Figure 2.1: Diagram of Artificial neural network	12
Figure 2.2: Bagging algorithm	22
Figure 2.3: Random forest algorithm.....	24
Figure 3.1: Pseudocode of the Levenberg-Marquardt algorithm	38
Figure 3.2: Effect of rule prioritization	43
Figure 3.3: Architecture of the proposed ensemble model	44
Figure 4.1: Test setup.....	48
Figure 4.2: GUI interface for entering validation rules.....	50
Figure 4.3: Accuracy of the model and its base classifiers	56
Figure 4.4: Mean processing times	59
Figure 4.5: Comparative accuracy of FDM model with other algorithms	61
Figure 4.6: Mean processing times of FDM model and other algorithms	65
Figure 7.1: Implementation of RIPPER algorithm.....	73
Figure 7.2: Request message from Switch.....	74
Figure 7.3: Response message to Switch	75

List of Tables

Table 2.1: Comparison of approaches	20
Table 3.1: Attributes used in transactions	34
Table 3.2: Rule prioritization - Influence on efficiency	42
Table 4.1: Details of the test datasets	47
Table 4.2: Results of Dataset 1	51
Table 4.3: Results of Dataset 2	51
Table 4.4: Results of Dataset 3	51
Table 4.5: Confusion matrix of binary classification	52
Table 4.6: Results of Rule-based classification with Dataset 1	54
Table 4.7: Results of Neural network with Dataset 1	55
Table 4.8: Accuracy of the model and its base classifiers	55
Table 4.9: Efficiency of the algorithms	58
Table 4.10: Results of the C 4.5 algorithm with Dataset 1	60
Table 4.11: Results of the CART algorithm with Dataset 1	60
Table 4.12: Testing for significance of accuracy	63
Table 4.13: Comparison of efficiency of algorithms	64

List of Abbreviations

Abbreviation	Description
ANN	Artificial Neural Networks
ATM	Automated Teller Machine
API	Application Programming Interface
EMV	Europay, MasterCard, Visa Standard
FDM	Fraud Detection Model
NFC	Near Field Communication
PIN	Personal Identification Number
POS	Point-Of-Sale
RFID	Radio-Frequency Identification
SMS	Short Message Service
SVC	Stored Value Card (also known as Prepaid Cards)