School of Information Technology

Enhancing Classification Performance over Noise and Imbalanced Data Problems

Piyasak Jeatrakul

This thesis is presented for the Degree of

Doctor of Philosophy of

Murdoch University

March, 2012

Declaration

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

Piyasak Jeatrakul

March, 2012

Acknowledgments

As there are a number of supporters who helped me to complete this thesis, I would like to take this opportunity to thank and acknowledge the following people and organisations.

First of all, I would like to express my immense gratitude and admiration towards my supervisor, Associate Professor Dr. Kevin Wong, for his advice, inspiration, and continuous encouragement throughout the course of the PhD program. He has not only offered me valuable recommendation on my work but also provided me great advice on my private issues. I am very grateful for his efforts and his assistance during the wonderful time at Murdoch University.

I would also like to give many thanks to my co-supervisor, Associate Professor Dr. Lance Chun Che Fung, for his helpful advice and mental stimulation at all times since I started the study four years ago. He has spent a lot of time sharing his research experiences and demonstrated his enthusiastic attitude towards research in several of his seminars and group meetings. Thanks again go to him for his useful comments on my work.

I am grateful to my PhD colleagues and Thai friends, who provided me friendship, care and assistance during the time I studied in Perth. The support of my parents, my wife and my children is the enormous power which has motivated me to complete my study.

Finally, without the support from Royal Thai Government and Mae Fah Luang University by providing financial support and the opportunity to study in Australia, it would have been impossible for me to accomplish the PhD program.

Abstract

This research presents the development of techniques to handle two issues in data classification: *noise* and *imbalanced data problems*. Noise is a significant problem that can degrade the quality of training data in any learning algorithm. Learning algorithms trained by noisy instances generally increase misclassification when they perform classification. As a result, the classification performance tends to decrease. Meanwhile, the imbalanced data problem is another problem affecting the performance of learning algorithms. If some classes have a much larger number of instances than the others, the learning algorithms tend to be dominated by the features of the majority classes, and the features of the minority classes are difficult to recognise. As a result, the classification performance of the minority classes could be significantly lower than that of the majority classes. It is therefore important to implement techniques to better handle the negative effects of noise and imbalanced data problems.

Although there are several approaches attempting to handle noise and imbalanced data problems, shortcomings of the available approaches still exist. For the noise handling techniques, even though the noise tolerant approach does not require any data pre-processing, it can tolerate only a certain amount of noise. The classifier developed from noisy data tends to be less predictive if the training data contains a great number of noise instances. Furthermore, for the noise elimination approach, although it can be easily applied to various problem domains, it could degrade the quality of training data if it cannot distinguish between noise and rare cases (exceptions). Besides, for the imbalanced data problem, the available techniques used still present some limitations. For example, the algorithm-level approach can perform effectively only on specific

problem domains or specific learning algorithms. The data-level approach can either eliminate necessary information from the training set or produce the over-fitting problem over the minority class. Moreover, when the imbalanced data problem becomes more complex, such as for the case of multi-class classification, it is difficult to apply the re-sampling techniques (the data-level approach), which perform effectively for imbalanced data problems in binary classification, to the multi-class classification. Due to the limitations above, these lead to the motivation of this research to propose and investigate techniques to handle noise and imbalanced data problems more effectively.

This thesis has developed three new techniques to overcome the identified problems. Firstly, a cleaning technique called the Complementary Neural Network (CMTNN) data cleaning technique has been developed in order to remove noise (misclassification data) from the training set. The results show that the new noise detection and removal technique can eliminate noise with confidence. Furthermore, the CMTNN cleaning technique can increase the classification accuracy across different learning algorithms, which are Artificial Neural Network (ANN), Support Vector Machine (SVM), *k*-Nearest Neighbor (*k*-NN), and Decision Tree (DT). It can provide higher classification performance than other cleaning methods such as Tomek links, the majority voting filtering, and the consensus voting filtering.

Secondly, the CMTNN re-sampling technique, which is a new under-sampling technique, has been developed to handle the imbalanced data problem in binary classification. The results show that the combined techniques of the CMTNN re-sampling technique and Synthetic Minority Over-sampling Technique (SMOTE) can perform effectively by improving the classification performance of the minority class

instances in terms of Geometric Mean (G-Mean) and the area under the Receiver Operating Characteristic (ROC) curve. It generally provides higher performance than other re-sampling techniques such as Tomek links, Wilson's Edited Nearest Neighbor Rule (ENN), SMOTE, the combined technique of SMOTE and ENN, and the combined technique of SMOTE and Tomek links.

For the third proposed technique, an algorithm named One-Against-All with Data Balancing (OAA-DB) has been developed in order to deal with the imbalanced data problem in multi-class classification. It can be asserted that this algorithm not only improves the performance for the minority class but it also maintains the overall accuracy, which is normally reduced by other techniques. The OAA-DB algorithm can increase the performance in terms of the classification accuracy and F-measure when compared to other multi-class classification approaches including One-Against-All (OAA), One-Against-One (OAO), All and One (A&O), and One Against Higher Order (OAHO) approaches. Furthermore, this algorithm has shown that the re-sampling technique is not only used effectively for the class imbalance problem in binary classification but it has been also applied successfully to the imbalanced data problem in multi-class classification.

List of Publications Related to this Thesis

The following six publications reported the results during the course of this research.

Journal

P1. P. Jeatrakul, K. W. Wong, and C. C. Fung, "Data cleaning for classification using misclassification analysis," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 3, pp. 297-302, 2010.

Lecture Notes in Computer Science

P2. P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," *Lecture Notes in Computer Science*, Springer Verlag, LNCS 6444, Issue Part 2, 2010, pp. 152-159.

Conference Proceedings

P3. P. Jeatrakul, K. W. Wong, C. C. Fung, and Y. Takama, "Misclassification analysis for the class imbalance problem," in *World Automation Congress (WAC 2010)*, Kobe, Japan, 2010, pp. 1-6.

P4. P. Jeatrakul, K. W. Wong, and C. C. Fung, "Using misclassification analysis for data cleaning," in *Proceedings of International Workshop on Advanced*

Computational Intelligence and Intelligent Informatics (IWACIII 2009), Tokyo, Japan, 2009, pp. PM-11.

P5. P. Jeatrakul and K. W. Wong, "Enhance the performance of complementary neural network using misclassification analysis," in *Proceedings of the Tenth Postgraduate Electrical Engineering and Computing Symposium (PEECS 2009)*, Perth, Australia, 2009.

P6. P. Jeatrakul and K. W. Wong, "Comparing the performance of different neural networks for binary classification problems," in *Proceedings of the Eighth International Symposium on Natural Language Processing (SNLP 2009)* Bangkok, Thailand, 2009, pp. 111-115.

Contributions of this Thesis

The contributions in this thesis, which have already been published and reported, can be described below and summarised in Table 1.

A review of different types of noise and imbalanced data, and a survey of various techniques to handle noise and imbalanced data problems have been completed. The features of classification techniques related to the research have also been reviewed. This work forms the basis of Chapter 2. Different parts of the work have been published in papers P1-6.

The development of the cleaning technique called the CMTNN cleaning technique forms a part of Chapter 3. The progress of the work, which includes algorithms, experimental studies, comparison results, and discussions, has been reported in journal paper P1 and conference papers P4 and P5.

The contribution in Chapter 4 is the successful development of data balancing algorithms to handle imbalanced data problems in binary classification. Several proposed re-sampling techniques have been explored and compared. The results of this work have been published in lecture notes in computer science paper P2 and a conference paper P3.

An algorithm to handle the class imbalance problem in multi-class classification has been successfully developed in Chapter 5. The proposed algorithm is the integration of the data balancing algorithm in paper P2 and a conventional multi-binary classification

technique named One-Against-All (OAA). The experimental results in this chapter have shown the significant improvement in terms of the classification performance after the algorithm has been implemented.

Table 1: Summary of the Contribution of the Thesis

Chapter	Contributions	Paper No
***Chapter 2*: Background**	Presents a literature survey on previous research related to noise and imbalanced data problems, and classification techniques used in the research.	P1, P2, P3, P4, P5, P6
***Chapter 3*: Noise Detection and Elimination Using CMTNN Cleaning Technique**	Successfully developed a technique for data cleaning by using misclassification analysis to eliminate noise with confidence.	P1, P4, P5
***Chapter 4*: CMTNN Re-Sampling Technique for Class Imbalance Problems**	Successfully developed a data balancing algorithm to handle imbalanced data problems in binary classification.	P2, P3
***Chapter 5*: Handling the Class Imbalance Problem in Multi-Class Classification**	Successfully enhanced the performance of the multi-class classification with imbalanced data problems by integrating the data balancing algorithm and a conventional multi-binary classification technique.	P2

Contents

List of Figures

List of Tables

List of Abbreviations

A&O	All and One
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
BNs	Bayesian Networks
BPNN	Back-Propagation Neural Networks
CMTNN	Complementary Neural Network
CNN	Condensed Nearest Neighbor Rule
DT	Decision Tree
ENN	Wilson's Edited Nearest Neighbor Rule
F_1	F-Measure
Falsity NN	Falsity Neural Network
FN	False Negative
FP	False Positive
FPSVM	Fuzzy Proximal Support Vector Machine
GEPSVM	Generalised Eigenvalue and Proximal Support Vector Machine
G-Mean	Geometric Mean
k-NN	*k*-Nearest Neighbor
LS-SVM	Least Squares SVM
ML	Machine Learning
NCL	Neighborhood Cleaning Rule
OAA	One-Against-All
OAA-DB	One-Against-All with Data Balancing
OAHO	One Against Higher Order
OAO	One-Against-One
OSS	One-Sided Selection
PF	Partitioning Filter
PNN	Probabilistic Neural Network
PRMs	Probabilistic Relational Models
PRMs-IM	Probabilistic Relational Models for Imbalanced Relational Data

PSVM	Proximal Support Vector Machine
RBFNN	Radial Basis Function Neural Network
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SPECT	Single Proton Emission Computed Tomography
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
Truth NN	Truth Neural Network
UCI	University of California Irvine