LOGISTIC REGRESSION METHODS FOR CLASSIFICATION OF IMBALANCED DATA SETS

SANTI PUTERI RAHAYU

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Doctor of Philosophy in Computer Science

Faculty of Computer System and Software Engineering
UNIVERSITI MALAYSIA PAHANG

SEPTEMBER 2012

ACKNOWLEDGEMENTS

All praise and thanks are Allah's, the Lord of the 'Alamin (mankind, jinn and all that exist). Alhamdulillah, thanks God, to make this research possible.

I would to express my deep gratitude to Prof. Dr. Jasni Mohamad Zain for the freedom to determine the path of my PhD Research and for having served as my advisor during my study in University Malaysia Pahang (UMP), Malaysia. I would also like to thank Prof. Dr. Abdullah Embong for his advice and wisdom.

I would like to thank Prof. Sabira K., Prof. Dr. Siti Mariyam binti Shamsuddin and Dr. Tutut Herawan for their valuable input, comments and suggestions. My appreciation also goes to all my 'teachers', either formal or informal.

I would like to acknowledge UMP for giving me a financial support during my study. I would also like to acknowledge Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS) Surabaya (Indonesia) for giving me a chance to further study in UMP.

I would like to thank Dr. Juwari Purwo Sutikno for his patience and invaluable help specially in providing basic of Matlab programming in this research and in editing layout of this thesis. I would like also to thank Dr. Anwaruddin Hisyam for taking time to edit English in this thesis. My appreciation also goes to all researchers for their useful researches which are the references of this research.

I would like to thank my colleagues (specially for Santi Wulan Purnami and Wibawati) and my friends (specially for Sita Fitriana, Dewi Anggoro, Amik Purbowati, Rini Susanah, Mulyati Ajisman, Wanti Utami, Ambikka and Mamiek Setyaningsih) for their attention, invaluable support and help during my study.

Finally, I would like to give my warm thanks to my husband, my children, my parents, my parents in law, my grandparents, my brothers-sisters and my big family for their love, wisdom, understanding, patience and strong support during my study.

ABSTRACT

Classification of imbalanced data sets is one of the important researches in Data Mining community, since the data sets in many real-world problems mostly are imbalanced class distribution. This thesis aims to develop the simple and effective imbalanced classification algorithms by previously improving the algorithms performance of general classifiers i.e. Kernel Logistic Regression Newton-Raphson (KLR-NR) and Regularized Logistic Regression NR (RLR-NR) which are Logistic Regression (LR)-based methods. Both LR-based methods have strong statistical foundation and well known classifiers which have simple solution of unconstrained optimization problem in performing the good performance as well as Support Vector Machine (SVM) which is determined as state-of-the art classifier in Kernel methodology and Data Mining community. However, the imbalanced LR-based methods are not extensively developed such as imbalanced SVM-based methods. Hence, it is required to develop effective imbalanced LR-based methods to be widely used in data mining applications.

Numerical results have showed that the use of Truncated Newton method for KLR-NR and RLR-NR which respectively resulted in Newton Truncated Regularized KLR (NTR-KLR) and NTR RLR (NTR-LR), is effective in handling the numerical problems on the huge matrix of *linear system of Newton-Raphson update rule* i.e. the training time and the singularity problem. These results can be seen as further explanation on the success of Truncated Newton method in TR-KLR and TR Iteratively Re-weighted Least Square (TR-IRLS) algorithm respectively, because of the equivalence of iterative method used by these algorithms. Moreover, only with the use of simple solution of unconstrained optimization problem, numerical results have demonstrated that proposed NTR-KLR and proposed NTR-LR respectively have comparable classification performance with RBFSVM (SVM with Radial Basis Function Kernel).

The imbalanced problem of both proposed general classification algorithms which is the limitation of accuracy performance specifically in classifying on the minority class has motivated this research to improve their classification performance on imbalanced data sets. In general, numerical results have showed that the use of adapted Modified AdaBoost methods for NTR-KLR and NTR-LR which respectively resulted in AdaBoost NTR Weighted KLR (AB-WKLR) and AB NTR Weighted RLR (AB-WLR) is significantly successful in improving the accuracy and stability performance of general classifiers i.e. NTR-KLR and NTR-LR respectively. The improvements on both error by *g-means* and *standard deviation* of *g-means* with *5*-Fold SCV could be achieved as high as more than 60. Furthermore, numerical results have demonstrated that proposed AB-WKLR and proposed AB-WLR respectively have comparable performances with AdaBoostSVM in classifying imbalanced data sets, only with the use of simple solution of unconstrained weighted optimization problem. Thus, both proposed imbalanced LR-based methods is simple and effective for classification of imbalanced data sets and have promising results.

ABSTRAK

Pengelasan set data yang tidak seimbang adalah salah satu kajian yang penting dalam masyarakat perlombongan data, kerana set data yang digunakan dalam dunia sebenar kebanyakannya adalah pengagihan kelas tidak seimbang. Tesis ini bertujuan untuk membangunkan algoritma pengelasan tidak seimbang yang mudah dan berkesan dengan meningkatkan prestasi algoritma pengelas umum iaitu *Kernel Logistic Regression Newton-Raphson* (*KLR-NR*) dan *Regularized Logistic Regression NR* (*RLR-NR*) yang merupakan kaedah berasaskan *Logistic Regression* (*LR*). Kedua-dua *LR-based methods* mempunyai asas statistik yang kukuh dan terkenal sebagai pengelas yang mempunyai penyelesaian yang mudah dari *unconstrained optimization problem* dalam melaksanakan prestasi yang sama baik dengan *Support Vector Machine* (*SVM*) yang ditentukan sebagai *state-of-the-art* pengelas dalam metodologi Kernel dan masyarakat Perlombongan Data. Walau bagaimanapun, *imbalanced LR-based methods* tidak dibangunkan secara meluas seperti *imbalanced SVM-based methods*. Oleh itu, ia diperlukan untuk membangunkan *imbalanced LR-based methods* yang berkesan yang digunakan secara meluas dalam banyak aplikasi perlombongan data.

Keputusan berangka telah menunjukkan bahawa penggunaan kaedah *Truncated Newton* untuk *KLR-NR* dan *RLR-NR* yang masing-masing mengakibatkan *Newton Truncated Regularized KLR* (*NTR-KLR*) dan *NTR RLR* (*NTR-LR*), adalah berkesan dalam menangani masalah berangka pada matriks besar dari sistem linear *Newton-Raphson update rule* iaitu masalah masa latihan dan ketunggalan. Keputusan ini boleh dilihat sebagai penjelasan lanjut mengenai kejayaan kaedah *Truncated Newton* di *TR-KLR* dan *TR Iterative Re-weighted Least Square* (*TR-IRLS*) algoritma, kerana kesetaraan kaedah lelaran yang digunakan oleh algoritma-algoritma ini. Selain itu, dengan hanya menggunakan penyelesaian yang mudah dari *unconstrained optimization problem*, keputusan berangka telah menunjukkan bahawa cadangan *NTR-KLR* dan cadangan *NTR-LR* masing-masing mempunyai prestasi klasifikasi setanding dengan *RBFSVM* (*SVM* dengan *Radial Basis Function*).

Masalah tidak seimbang kedua-dua algoritma klasifikasi umum yang dicadangkan yang merupakan had prestasi ketepatan khususnya dalam mengklasifikasikan kelas minoriti telah mendorong kajian ini untuk meningkatkan prestasi klasifikasi mereka pada set data yang tidak seimbang. Secara umum, keputusan berangka telah menunjukkan bahawa penggunaan kaedah *adapted Modified AdaBoost* untuk *NTR-KLR* dan *NTR-LR* yang masing-masing mengakibatkan *AdaBoost NTR Weighted KLR* (*AB-WKLR*) dan *AB NTR Weighted RLR* (*AB-WLR*) adalah lebih berjaya dalam meningkatkan prestasi ketepatan dan kestabilan pengelas umum iaitu *NTR-KLR* dan *NTR-LR*. Peningkatan bermakna oleh kedua-duanya atas kesilapan *g-means* dan sisihan piawai *g-means* dengan 5-Lipat SCV boleh dicapai setinggi lebih daripada 60. Tambahan pula, keputusan berangka telah menunjukkan bahawa cadangan *AB-WKLR* dan cadangan *AB-WLR* masing-masing mempunyai persembahan yang setanding dengan *AdaBoostSVM* dalam mengklasifikasikan set data tidak seimbang, hanya dengan menggunakan penyelesaian yang mudah dari *unconstrained weighted optimization problem*. Oleh itu, kedua-dua cadangan imbalanced *LR-based methods* merupakan kaedah yang mudah dan berkesan untuk pengkelasan set data yang tidak seimbang dan mendapat keputusan yang menjanjikan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

a	The optimal step length
α	Coefficient vector of Kernel Logistic Regression
β	Coefficient vector of Regularized Logistic Regression
c	Conjugacy enforcer
d	The search direction
dis	The distance of a sample from the separating hyperplane
dim	Number of attributes
D	Diagonal matrix of variance and weight vector
ε_1	Threshold of the difference of optimization function values
ε_2	The convergence threshold for Linear CG
ε_t	The weighted error of component classifier on *t*-th round
f	Linear function
F	Ensemble function
h_t	The weighted prediction of component classifier on *t*-th round
g	Gradient vector
H	Hessian matrix
K	Kernel matrix
k	Number of fold
k_{ij}	Cell of kernel matrix
$\mathbf{K}_1$	Kernel matrix with the bias term
$\mathbf{K}_2$	Matrix that consist of diagonal element: **K** and the bias term is not regularized
l	Likelihood function

L	Log-likelihood function
λ	Regularization parameter
n	Number of samples
n_σ	Number of σ during AB-WKLR iterations
n_λ	Number of λ during AB-WLR iterations
$\mathbf{p}$	Probability of given input
φ	Function to map the original data $\mathbf{x}$ in input space into feature space
q	Quadratic form
$\mathbf{r}$	Residual
$\mathbf{s}$	Vector of Newton direction
S_{gmeans}	Standard deviation of g-means values
σ	RBF Kernel parameter
T	the number of AdaBoost iterations,
$\boldsymbol{\theta}$	Vector of general parameter
$\mathbf{v}$	Variance vector
$\mathbf{V}$	Diagonal matrix of $\mathbf{v}$
$\mathbf{w}$	Weight vector of training samples
$\mathbf{W}$	Diagonal matrix of $\mathbf{w}$
ω	The importance factor of corresponding component classifier to an ensemble
$\mathbf{x}$	Input vector without bias term
$\mathbf{y}$	Vector of input label
y_{pred}	Predictions of AdaBoost classifier
Z_t	The normalization factor on *t*-th round
$\mathbf{Z}$	Vector of adjusted response

ζ	Function of Bernoulli distribution

LIST OF ABBREVIATIONS

AB-WKLR	Adaptive Boosting Weighted Kernel Logistic Regression
AB-WLR	Adaptive Boosting Weighted Regularized Logistic Regression
AdaBoost	Adaptive Boosting
AdaBoostSVM	Adaptive Boosting Support Vector Machine
AUC	Area Under Receiver Operating Curve
CG	Conjugate Gradient
CM	Confusion Matrix
CV	Cross Validation
DEV	Deviance
IVM	Import Vector Machine
NR	Newton-Raphson
NRUR	Newton-Raphson update rule
NTR-LR	Newton Truncated Regularized Logistic Regression
NTR-KLR	Newton Truncated Regularized Kernel Logistic Regression
NTR-WKLR	Newton Truncated Regularized Weighted Kernel Logistic Regression
NTR-WLR	Newton Truncated Regularized Weighted Regularized Logistic Regression
GS	Grid Search
GSVM-RU	Granular Support Vector Machine-Repetitive Under-sampling
IRLS	Iteratively Re-Weighted Least Square
KLR	Kernel Logistic Regression
KLR-IRLS	Kernel Logistic Regression Iteratively Re-Weighted Least Square

KLR-NR	Kernel Logistic Regression Newton-Raphson
LCG	Linear Conjugate Gradient
MLE	Maximum Likelihood Estimation
NLL	Negative Log-Likelihood
NR	Newton-Raphson
RBF	Radial Basis Function
RE-WKLR	Rare Event Weighted Kernel Logistic Regression
RLR	Regularized Logistic Regression
RLR-IRLS	Regularized Logistic Regression Iteratively Re-Weighted Least Square
RLR-NR	Regularized Logistic Regression Newton-Raphson
SCV	Stratified Cross Validation
SDC	Smote with Different Cost
SMO	Sequential Minimization Organization
SMOTE	Synthetically Minority Over-sampling Technique
SVM	Support Vector Machine
TR-IRLS	Truncated Regularized Iteratively Re-Weighted Least Square
TR-KLR	Truncated Regularized Kernel Logistic Regression
WKLR	Weighted Kernel Logistic Regression
WLR	Weighted Regularized Logistic Regression
WLS	Weighted Least Square
WWBOOST-SVM	Weighting rule and Weakened Support Vector Machine based Boosting

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The interface of statistics, database technology, pattern recognition, machine learning, and other areas are termed as Data Mining. It is concerned with the analysis of large databases by using machine learning methods, in identifying previously unsuspected pattern which are of interest or value to the data. (Hand, 1998; Tan et al., 2005).

Classification is a supervised data mining task, which is a predictive task with qualitative outcome. In the last decade, it is found that, beside the evaluation of data in manual, the use of classifier system is also very important factor in helping expert to make decision, i.e. to identify pattern and make prediction. Classifier system can achieve a fast, objective, more detailed and accurate classification by minimizing possible errors due to fatigued or inexperienced expert. (Huang et al., 2007; Polat et al., 2007; West, 2000).

In the last decade, the resulting family of Kernel learning methods (Scholkopf and Smola, 2002; Shawe and Christianini, 2004) have frequently demonstrated state-of-the-art performance on a wide range of benchmark and real-world applications. Most of these kernel-based methods, however, are presented in the literature along with the Support Vector Machine (SVM) method. SVM (Vapnik, 1998; Vapnik, 2000), which was developed based on the theory of Structural Risk Minimization (SRM), is popular with its effectiveness in the Kernel Machine Learning and Data Mining Community,

such that it is considered as state-of-the-art algorithm for classifying non-linear binary data.

Beside SVM, Kernel Logistic Regression (KLR) (Roth, 2001; Zhu and Hastie, 2004; Zhu and Hastie, 2005) is one of the most important recent developments for classification task in Kernel-machine techniques. It is the Kernel version of Regularized Logistic Regression (RLR) (Minka, 2003; Zhang and Oles, 2001) classifier. The use of Kernel in KLR algorithm is to improve the generalization performance of RLR on overcoming the non-linear problem that has low-to-medium-dimensional data (Maalouf, 2009).

Meanwhile, RLR is the regularized version of Logistic Regression (LR) (Hosmer and Lemeshow, 2000; Dreitsel and Machado, 2002; Hastie et al., 2001; McCulagh and Nelder, 1989) which is the fundamental and well known statistical method for classification task. It is a classifier which is well applied to linear problem with high-dimensional data (Komarek and Moore, 2005). Hence, RLR is considered as state-of-the-art algorithm for linear discriminant data.

KLR and RLR have received more extensive research attention, since they have similar loss function with SVM (Patra et. al., 2008; Rahimi, 2006; Rennie, 2005; Zhang and Oles, 2001; Zhang et al., 2003; Zhu and Hastie, 2005). Furthermore, by using *total accuracy* metric, the classification performance of KLR is similar to non-linear SVM (Karsmaker et al., 2007), while the classification performance of RLR is comparably accurate to linear SVM (Zhang et al., 2003; Zhang and Oles, 2001). However, optimization of SVM needs to be solved with quadratic constrained optimization, while KLR and RLR only need to be solved by unconstrained optimization (Maalouf, 2009), although it also can be stated as constrained optimization problem (Karsmaker et al., 2007; Kerthi et al., 2005). In addition, unlike SVM, both classifiers naturally provide probability of classification membership (Zhu, 2003; Zhang et al., 2003).

Many problem domains require transparent reasoning as well as accurate classifier (Ridgeway et.al, 1998). Trust in a system is developed by the quality of the results (accuracy) and also by clear description of how they were derived (transparent

reasoning) (Swartout, 1983). Good accuracy enables correct assessments / diagnosis / treatment and thus avoiding any heavy losses associated with wrong prediction (Lahsasna et al, 2008; West, 2000). Transparency enables expert to understand the classification/decision process. The capability of classifier to describe its analysis often affects the end-user acceptance. In types of situation like these, LR-based methods, i.e. KLR and RLR, are appropriate methods.

In summary, LR-based methods have simple optimization function than SVM-based methods on performing comparable accuracy. Moreover, the transparency of LR-based methods is supported by providing the membership probability naturally. Furthermore, LR-based methods are well known methods and have strong statistical foundation. However, as further as limited knowledge, the LR-based methods have less extensive research than SVM-based methods on imbalanced classification problem. Hence, in order to take the advantages of LR-based methods and to give further contribution on the research of LR-based methods, this thesis aims to further develop the LR-based methods for solving the classification problems, either general or imbalanced problem.

1.2 PROBLEM STATEMENT AND MOTIVATION

This thesis interests to conduct study on two main problems of KLR and RLR. The problems can be stated as follows:

(i) Newton-Raphson (Rennie, 2003) is the most commonly method to solve the non-linear optimization problem of KLR and RLR. Newton-Raphson method iteratively solves the linear system of Newton-Raphson Update Rule (NRUR). As has been reported in literatures, however, the use of Newton-Raphson method for KLR and RLR has numerical problem that the huge Hessian matrix needs to be inverted (Lin et al., 2008; Zhu and Hastie, 2005). Due to the density of its matrices, their computation can be slow (Komarek, 2004; Karsmakers et al. 2007; Maalouf, 2009).

(ii) General classifiers, such as SVM, KLR and RLR, were developed and evaluated on the assumption that the data has balanced class distribution (Japkowicz,

2000; Maalouf, 2009). However, in many real-world problems, it was faced that the data sets have imbalanced class distribution. The class imbalance problem corresponds to domains for which one class is represented by a large number of examples while the other is represented by only a few (Guo and Viktor, 2004; Japkowicz, 2000). In the case of binary classification, data sets are said to be imbalanced, if the number of negative instances are heavily larger than the positive ones (Akbani et al., 2004; Maalouf, 2009). Commonly, for two-class classification of imbalanced data set, the negative class is the notation for the majority class, while the positive class is the notation for the minority class. In imbalanced classification problems, the minority class is the class of primary interest. As has been reported in literatures of Kernel learning, it seems difficult for general classifier algorithms, even though SVM, to detect regularities within the minority class on imbalanced data problems (Akbani et al, 2004; Maalouf, 2009). Therefore, they have good specificity, but poor sensitivity (Akbani et al., 2004; Maloouf, 2003). King and Zeng (2001c) stated similarly that when non-kernel of probabilistic method such as logistic regression, is used, it underestimates the probability of rare events, because it tends to be biased towards the majority class, which is the less important class. Recently, in relation to further development of KLR and RLR respectively, this thesis has confirmed the limitation performance of both general classification algorithms on imbalanced data sets. The report can be found in Chapter 4.

The motivation of this research is described as follows:

(i) Several methods have been proposed for solving the numerical problem of KLR and RLR. Detail analysis of those methods proposed will be reported in Chapter 2. In the last decade, the use of Truncated Newton methods are the most proposed methods on applying KLR and RLR. However, so far, the success of Truncated Newton method in both algorithms has not been totally explored. Therefore, this thesis intends to contribute further explanation on the success of Truncated Newton for KLR and RLR specifically on improving the algorithm performance of these both LR-based methods.

(ii) For solving the imbalanced classification problem, a number of methods have been proposed in literatures of Kernel learning. Discussion on the limitation of those methods will be reported in detail, in Chapter 2. Based on those methods proposed, in general, the research of imbalanced LR-based methods are not as many as the research of imbalanced SVM-based methods which have good accuracy performance. Furthermore, the imbalanced techniques used on LR-based methods have led their accuracy performances for classification of imbalanced data sets that still require an improvement. Hence, it is important to develop the effective imbalanced LR-based methods for solving the imbalanced classification problem of general LR-based methods.

1.3 THE APPROACHES

This research concerns on developing better general and imbalanced classification algorithms for KLR-NR and RLR-NR. Related to this concern, there are two main problems that must be handled in this thesis, as stated in the previous section. The approach for solving those problems can be described as follows:

(i) In order to develop the simple and effective of general classification algorithms for KLR-NR and RLR-NR respectively, this research proposes the implementation of Truncated Newton method. Among other Truncated Newton LR-based method, the simplicity and the effectiveness of Truncated Regularized KLR (TR-KLR) (Maalouf et al., 2010) and TR Iteratively Re-weighted Least Square (TR-IRLS) (Komarek and Moore, 2005) have inspired this research. TR-KLR is as accurate as, and much faster than, non-Linear SVM on small-to-medium size data sets of non-linear classification problem. Meanwhile, TR-IRLS is comparably accurate with, and faster than, Linear SVM on large size data sets of linear classification problem.

In general, the use of Truncated Newton method typically consists of truncated inner algorithm and outer algorithm (Nash, 2000). In TR-KLR and TR-IRLS, the use of Truncated Newton includes Linear Conjugate Gradient (CG) method (Gilbert, 2006; Nash and Sofer, 1996; Shewchuk, 1994) and Iteratively Re-

weighted Least Square (IRLS) procedure (Mc Cullagh and Nelder, 1989; Nabney, 1999; Hastie et al., 2001) for KLR and RLR respectively.

In summary, the approaches for solving the numerical problem of KLR-NR and RLR-NR can be explained as follows:

(a) It is necessary to keep the use of unconstrained optimization problem for KLR-NR and RLR-NR respectively. This optimization problem typically has simpler solution than the constrained ones.

(b) It is also necessary to keep the use of Linear CG method, as the truncated inner algorithm of Truncated Newton method for KLR and RLR respectively. This method has faster computation in approximating the Newton's solution.

(c) Instead of IRLS procedure as used by TR-KLR and TR-IRLS, this approach uses Newton-Raphson method as the outer algorithm of Truncated Newton method. Newton-Raphson and IRLS are equivalent method for KLR and RLR. In addition, Newton-Raphson method is mathematically simple, because IRLS procedure is a representation of Newton-Raphson method.

The use of Truncated Newton method for solving the numerical problem of KLR-NR and RLR-NR algorithm respectively results in proposed Newton TR-KLR (NTR-KLR) and proposed Newton TR RLR (NTR-LR) algorithm. Because of the equivalency between Newton-Raphson method and IRLS procedure, the accuracy performance of both proposed classifier can be expected to have similar performance for TR-KLR and TR-IRLS respectively. In addition, both proposed algorithms can be seen as the Newton version of TR-KLR and the Newton version of TR-IRLS algorithm. Hence, both proposed algorithms can be used to contribute further explanation on the success of Truncated Newton method in TR-KLR and TR-IRLS respectively.

Moreover, the development of both proposed algorithms can be seen as preliminary representation of idea stated by Komarek (2004) that whether the behaviour of Newton-Raphson and Linear CG combination would be identical to IRLS and Linear CG combination. In specific, development of proposed NTR-KLR algorithm can be seen also as preliminary representation of Kernel version to the Trust Region Newton RLR that was proposed by Lin et al. (2008).

(ii) In order to develop the effective imbalanced classification algorithms for NTR-KLR and NTR-LR respectively, this thesis proposes the use of Modified AdaBoost method (with some adaptations). This is motivated by the success of imbalanced SVM-based method i.e. Adaptive Boosting SVM (AdaBoostSVM) (Li et al., 2008) with the use of this imbalanced technique. AdaBoostSVM has much better performance than SVM on solving the imbalanced classification problem. The use of AdaBoost-based method (Freund and Schapire, 1997) typically contains ensemble method and component classifier. In AdaBoostSVM, the ensemble method used is Modified AdaBoost and the component classifier is SVM with Radial Basis Function (RBF) Kernel (RBFSVM).

Detail strategies for solving the imbalanced classification problem of general LR-based methods are described in the following:

a. It is necessary to keep the use of Modified AdaBoost (with some adaptations) as the ensemble method of proposed imbalanced LR-based methods. Boosting mechanism of Modified AdaBoost forces the component classifiers to focus on the misclassified samples from the minority class by increasing the weights of training data. This prevents the minority class from being consider as noise in the majority class and be wrongly classified on imbalanced problem.
b. Instead of SVM, this approach uses NTR-KLR and NTR-LR respectively as the component classifier of proposed imbalanced LR-based methods. As proposed previously, NTR-KLR and NTR-LR are representation of KLR-NR and RLR-NR with Truncated Newton method respectively. The similarity of loss function among NTR-KLR, NTR-LR and SVM, has led these classifiers can be expected to have comparable accuracy. In addition, with the use of unconstrained optimization problem, NTR-KLR and NTR-LR have simpler solution of optimization problem than SVM.

The implementation of adapted Modified AdaBoost ensemble method for solving the imbalanced classification problem of NTR-KLR and NTR-LR component classifier respectively are called as Adaptive Boosting NTR Weighted KLR (AB-WKLR) and AB NTR Weighted RLR (AB-WLR) algorithm. As further as limited

knowledge, Nishida and Kurita (2006) were the first researchers who applied Boosting method, i.e. LogitBoost, on sparse version of KLR, i.e. Import Vector Machine (IVM) (Zhu and Hastie, 2005), While Huang et al. (2005) was the first to employ classic AdaBoost method on Logistic Regression (LR) that used weighted least-squares as the objective function and batch gradient descent algorithm for its optimization.

Since there is similarity loss function between component classifiers used, the accuracy performance of the proposed algorithms can be expected as well as AdaBoostSVM in classifying the imbalanced data sets. Moreover, the comparable accuracy only requires to be obtained by the simple solution of unconstrained optimization problem.

1.4 OBJECTIVES AND SCOPES

The main objective of the research is to develop the simple and effective classification algorithms using LR-based methods.
The research objective can be stated in detail as follows:

(ii) To develop general classification algorithms, i.e. NTR-KLR and NTR-LR
(iii) To develop imbalanced classification algorithms, i.e. AB-WKLR and AB-WLR

The scope of this research covers the following:

(i) This thesis considers 2-class classification and the data sets used mostly are imbalance.
(ii) Proposed general classification algorithms are developed based on KLR-NR and RLR-NR algorithm respectively, while proposed imbalanced classification algorithms were developed based on NTR-KLR and NTR-LR algorithm respectively.
(iii) Proposed NTR-KLR and proposed AB-WKLR are applied on small-to-medium size of data sets, while proposed NTR-LR and proposed AB-WLR are employed on large size data sets.

1.5 CONTRIBUTIONS

The primary contributions of this research are as follows:

(i) NTR-KLR and NTR-LR algorithm were developed. Both proposed algorithms contribute to the study of KLR-NR and RLR-NR respectively, by providing the simple and effective general classification algorithms for KLR-NR and RLR-NR respectively with the use of Truncated Newton method. Both proposed algorithms are also provided specifically to conduct further explanation on the success of Truncated Newton method in TR-KLR and TR-IRLS respectively, since both proposed algorithms are equivalent to TR-KLR and TR-IRLS respectively. In general, both proposed algorithms contribute to the general classification research of LR-based methods.

(ii) AB-WKLR and AB-WLR algorithm were developed. Both proposed algorithms contribute to the research of KLR-NR and RLR-NR with Truncated Newton method respectively, by providing the simple and effective imbalanced classification algorithms for NTR-KLR and NTR-LR respectively with the use of adapted Modified AdaBoost method. In general, both proposed algorithms contribute to the imbalanced classification research of LR-based methods.

1.6 OUTLINE OF THE THESIS

This thesis is organized as follows. Chapter 2 gives extended reviews of TR-IRLS, TR-KLR, AdaBoost algorithms for SVM and some basic theories of numerical experiment. Chapter 3 describes the proposed algorithms and the research methodology. In chapter 4, several numerical results of experiment are reported and discussed. At the end, conclusions for this research and recommendations for the further work are given in chapter 5.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter presents the reviews of General and Imbalanced Classification Research, including TR-IRLS, TR-KLR, Adaptive Boosting (AdaBoost) algorithms for SVM and some basic theories on conducting numerical experiment. These reviews are required as fundamental theory in order to propose new algorithm of KLR and RLR, on both the algorithmic level and in dealing with the imbalanced problems.

2.2 CLASSIFICATION

Globally, data mining tasks are divided into two categories, namely supervised and unsupervised task. As mentioned in Chapter 1, classification is a supervised data mining task on predicting categorical response.

In the last decade, there are many classification methods that have been proposed on general and imbalanced data assumption. Among other classification methods, the maturity of LR-based methods has motivated this thesis for exploring these methods as the simple and effective classifier to be widely used in data mining application, either on general or imbalanced data sets.

In order to develop better performance of general and imbalanced classification algorithms for LR-based methods i.e. KLR and RLR, it is important to study the limitation of related previous research. In the following, summary of the latest research of LR-based methods in relation with general and imbalanced data are reviewed. In