SUPPORT VECTOR MACHINE FOR SOLVING SMALL DATASET PROBLEM

AHMAD RIJAL BIN ABDUL RAHMAN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical – Mechatronics and Automatic Control)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JANUARY 2012

To my beloved mother Rofishah Binti Hj Zakaria and dedicated in memoirs to my late father Abdul Rahman Bin Lebai Ismail, whose don't have the opportunity to share my success. Al-Fatihah…

ACKNOWLEDGEMENT

In the Name of Allah, Most Gracious, Most Merciful. I am grateful to Allah for His guidance, and only by His strength I have successfully completed my master project and the write up on this thesis.

I would also wish to extend my gratitude and appreciation to my supervisor, Dr. Zuwairie Bin Ibrahim for his precious guidance, assistance, advice and positive comments throughout the accomplishment of this project. Appreciation and thankfulness to En Ibrahim Bin Shapiai and to all my friends for the encouragement, cooperation and inspiration they gave all along the way to the completion of this project.

Finally, I would like to thank my family for their determined support, encouragement and understanding. I am grateful to all these important peoples.

ABSTRACT

Data quantity is the main concern in the small data set problem, because usually insufficient data information will not lead to a robust classification performance. How to extract more effective information from a small data set is thus of considerable interest. A computational technique called Support Vector Machine (SVM) constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks, is proposed for this project. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin). In general, the larger the margin the lower the generalization error of the classifier is achieved. In this research, Support Vector Machine (SVM) is employed for solving small dataset problems in binary classification. A lot of performance measure can be used to measure the performance of data. This research used accuracy as a performance measure. In order to improve the performance of accuracy, SMOTE (Synthetic Minority Oversampling Technique) algorithm has been used to balance the data with creates a synthetic data in the minority class for imbalanced dataset or both of negative and positive class for balanced dataset problem. An algorithm of SVM and SMOTE has been developed using Matlab.

ABSTRAK

Kuantiti data adalah perkara utama yang perlu dititikberatkan dalam masalah set data yang kecil kerana pada kebiasaannya, kekurangan maklumat pada data tidak akan memberi ketepatan yang teguh dalam pengelasan data. Bagaimana untuk mengeluarkan maklumat yang lebih tepat dari set data yang kecil adalah perkara yang perlu dipertimbangkan. Projek ini telah mencadangkan teknik penaksiran yang di kenali sebagai *Support Vector Machine (SVM)* yang akan membentuk satu *hyperplane* atau set-set *hyperplane* dalam ruang dimensi yang luas atau ruang dimensi yang tidak terhingga yang mana boleh digunakan untuk pengelasan, regresi atau tugas-tugas yang lain. Tanpa perlu di persoalkan lagi, pemisahan yang baik telah dicapai oleh *hyperplane* yang mempunyai jarak terbesar dengan data latihan yang hampir dengan mana-mana kelas (dikenali sebagai *functional margin*). Dalam erti kata yang lain, semakin besar *margin* semakin kecil kesilapan umum oleh pengelas dicapai. Dalam penyelidikan ini, *Support Vector Machine (SVM)* di tugaskan untuk menyelesaikan masalah set data yang kecil dalam bentuk pengelasan perduaan. Dalam penyelidikan ini, ketepatan telah digunakan sebagai ukuran prestasi. Dalam usaha untuk meningkatkan tahap prestasi ketepatan, algoritme *SMOTE (Synthetic Minority Over-sampling Technique)* telah digunakan untuk mengimbangi data dengan membentuk satu data buatan dalam kelas yang terkecil untuk set data yang tidak seimbang atau untuk kedua-dua kelas positif dan negatif untuk set data yang seimbang. Algoritme SVM dan SMOTE telah dibina dengan menggunakan perisian Matlab.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
BPNN	-	Back Propagation Neural Network
BESVM	-	Boosting Evolutionary Support Vector Machine
BSVM	-	Biased Support Vector Machine
CSVM	-	Central Support Vector Machine
FMS	-	Flexible Manufacturing Scheduling
FN	-	False Negative
FP	-	False Positive
GA	-	Genetic Algorithm
KICA	-	Kernel Independent Component Analysis
KKT	-	Karush-Kuhn-Tucker
KPCA	-	Kernel Principle Component Analysis
MMC	-	Maximal Margin Classifier
MTD	-	Mega-Trend Diffusion
PCA	-	Principle Component Analysis
PPNN	-	Posterior Probability Neural Network
PSO	-	Particle Swarm Optimization
QP	-	Quadratic Programming
RBF	-	Radial Basic Function
SMOTE	-	Synthetic Minority Oversampling Technique
SVM	-	Support Vector Machine
TN	-	True Negative
TP	-	True Positive

LIST OF SYMBOLS

S	-	Training sample
L	-	Training set size
N	-	Dimensional input space
$H_{optimal}$	-	Optimal hyperplane
ξ	-	Slack variable
w	-	Weight vector
b	-	Bias
α	-	Dual variable
L	-	Primal lagrangian
W	-	Dual lagrangian
C	-	Margin parameter
K	-	Nearest neighbor parameter

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Small dataset conditions exist in many applications, such as disease diagnosis, fault diagnosis or deficiency detection in biology and biotechnology, mechanics, flexible manufacturing system scheduling, drug design, and short-term load forecasting (an activity conducted on a daily basis by electrical utilities). Neural networks have been applied successfully in many fields. However, satisfactory results can only be found under large sample conditions. When it comes to small training sets, the performance may not be so good, or the learning task can even not be accomplished. This deficiency limits the applications of neural network severely. Several computational intelligence techniques have been proposed to overcome the limits of learning from small datasets.

For this project a techniques that has been proposed is Support Vector Machine (SVM) as a classifier and the Synthetic Minority Over-Sampling Technique (SMOTE) as a data level. Support Vector Machine (SVM) classification is an active research area which solves classification problem in different domain. Support Vector Machine (SVM) is proposed by Vapnik *et al.* (2002) which used to find an optimal separating hyperplane. Support Vector Machine (SVM) is divided by four concepts: the separating hyperplane, the maximum-margin classifier, the soft margin and the kernel function. But this research only focus on separating hyperplane, the maximum-margin classifier and the soft margin. The hyperplane is used for the classification and used to separate the training data. A good separation can be achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin). The advantage of using Support Vector Machine (SVM) is SVM can prevent the overfitting training data by controlling the hyperplane margin measure. Optimization theory (quadratic programming) provides the mathematical techniques that necessary to find hyperplanes and optimize the measure. The Maximum-Margin Classifier (MMC) is the simplest model of Support Vector Machine (SVM) because it contain easiest algorithm to understand. But the MMC only works with data that linearly separable in the features space. Therefore the maximum-margin classifier cannot be used in many real world applications. In order to overcome this problem, the soft margin is introduced. Soft margin is the better way to solve the problem where the data are not linearly separable in features space (the algorithm of maximal margin and soft margin will be explained in chapter 3).

In the real world application, many techniques have been proposed as a data level approach such under sampling technique, over sampling technique and etc. For this project, Synthetic Minority Over-Sampling Technique (SMOTE) is proposed as the data level approach. Synthetic Minority Over-Sampling Technique (SMOTE) has been proposed by Nitesh V. Chawla *et al.* Basically, the SMOTE approach works when the minority class is over- sampled by creating a synthetic data. Nevertheless, this project only consists of balanced data. Therefore, this technique has been used to over-sample both positive and negative class (the algorithm of Synthetic Minority Over-Sampling Technique (SMOTE) will be explained in chapter 3).

1.1 Problem Statement

The main problem when involved with small datasets problem is the small datasets cannot provide good enough information. The main reason why small datasets cannot provide well enough information is the gaps between samples will be existed and the domain of samples cannot be ensured. Since the small dataset have not enough information, it will reduce the classification performance. The result also in the risk of over fitting of the training data and also can lead to poor generalization capabilities of the classifier.

1.2 Objectives

The goals of this project are:

i. To investigate a performances of accuracy by solving small and balance dataset problems by using Support Vector Machine.
ii. To investigate the changes of accuracy by using SMOTE (Synthetic Minority Oversampling Technique) algorithm in order to balance the data with create a synthetic data in the positive class and negative class.

1.3 Scope of Work

A scope of the project needs to be narrowed down, so it can be completed within two semesters. Following are the scope of this project:

i. An algorithm is developed using MATLAB^{TM} software.
ii. Used the small dataset problems and balanced datasets (Binary Classification).
iii. All datasets are taken from UCI machine learning.
iv. Used four types of datasets: Haberman’s Survival Dataset, Pima Indian Diabetes, German Credit and Liver Disorder.

v. Performance measure that has been used is accuracy
vi. Used Synthetic Minority Over-Sampling Technique (SMOTE) as a data level technique.

1.4 Thesis Overview

This thesis is organized into 5 chapters:

i. Chapter 1 : Introduction
ii. Chapter 2 : Literature Review
iii. Chapter 3 : Methodology
iv. Chapter 4 : Results and Discussions
v. Chapter 5 : Conclusion

Chapter 1 presents the introduction of the project. It included the overview of Support Vector machine (SVM) and Synthetic Minority Over-Sampling Technique (SMOTE). It also provides readers a first glimpse at the basic aspects of the research undertaken such as objectives, scope of work and problem statement.

Chapter 2 gives an insight to the research and development of Support Vector Machine and Synthetic Minority Over-Sampling Technique (SMOTE) in order to solve the small dataset problem and also detection done by various researchers and the background study of this project.

Chapter 3 presents the theories and methodology of the proposed method or technique. In this section, detailed explanation given for each stage involve in the development process.

Chapter 4 mainly devoted for demonstrating the experimental results of the project, performances of accuracy, analysis and discussions.

Chapter 5 presents the summary and conclusions of the project. Some recommendation and suggestions for the future development of the project are also discussed

1.5 Summary

In this chapter is, well planning is very important to make sure this project success. Every planning that planned should be follow to make this project finished on the dateline or earlier before the dateline. Besides that, this project has been developed based on the problem statements that are state in this chapter. The objective of the project is also important to make sure this project successfully and as aim of this project. In addition, the scope of the project needs to be recognizing before starting this project.

REFERENCES

Chawla, N.V., K.W. Boyer, K.W. and Kegelmege W.P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligent Research*. Vol. 16, 321-357.

Chawla, N.V., Boyer, K.W., Lazarevic, A. and Hall, L.O. (2003). SMOTEBoost: Improving Prediction of the Minority Class. *Proceeding of the Principle of Knowledge Discovery in Database.* PKDD-2003, 107-119.

DEEPA, T. and PUNITHAVALLI, M. (2011). An E-SMOTE Technique for Feature Selection in High-Dimensional Imbalanced Dataset. *Electronic Computer Technology (ICECT), 3rd International Conference.* 8-10 April, 322-324.

Der, C.L. and Chiao, W.L. (2010). Extending Attribute Information for Small Data Set Classification. *IEEE Transactions on Knowledge and Data Engineering*. Vol. PP (99). 30 December, 1.

Hui, L.H., Yi, H.C., Dwight, D.K. and Shinn, Y.H. (2007). Boosting Evolutionary Support Vector Machine for Designing Tumor Classifiers from Microarray Data. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. 1-5 April, 32-38.

Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Set: One Sided Selection. *Proceeding of the Fourteenth International Conference of Machime Learning*. Nashville, Tennessee: IEEE, 179-186.

Nello, C. and John, S.T. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, UK: Press Syndicate of The University of Cambridge.

Pero, R. and Srdjan, S. (2002). Neural Network Models Based on Small Data Sets. *6th Seminar on Neural Network Application in Electrical Engineering*.

September 26-28. Belgrade, Yugoslavia :IEEE, 101-106.

Razvan, A., Levente, F.A., Christopher, B., Abdul, W., Sarah, A.W., Grant I. B. and Lukas C. M. (2011). Fuzzy ARTMAP Prediction of Biological Activities for Potential HIV-1 Protease Inhibitors Using a Small Molecular Data Set. *ACM Transactions on Computational Biology and Bioinformatics*. Vol. 8(1), 80-93.

Rongfu, M.H.Z., Linke, Z.A.C. and Aizhi, C. (2006). A New Method to Assist Small Data Set Neural Network Learning. *Intelligent Systems Design and Applications (Sixth International Conference.*16-18 October. Jinan, 17-22.

S. Sivakumari, R. Praveena Priyadarsini and P. Amudha (2009). Performance Evaluation of SVM Kernels Using Hybrid PSO-SVM. *ICGST-AIML Journal, ISSN: 1687-4846*, Vol. 9 (1), 19-25.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. New York, USA: Springer-Verlag New York.

Vapnik, V.N. (1998). *Statistical Learning Theory*. New York, USA: John Wiley and Sons, New York, USA.

Wang, H.Y. (2008). Combination Approach of SMOTE and Biased-SVM for Imbalanced Dataset. *IEEE World Congress on Computational Intelligent*. 1-8 June, 22-21.

Wei, H.A.W., Ya, C.C. and Wen, H.C. (2010). A Research of Intelligent Parameters Searching in Small Data Sets. *Industrial Engineering and Engineering Management (IE & EM), 17th International Conference*. 29-31 October, 379-383.

Xuegong Zhang (1999). Using Class-Center Vectors to Build Support Vector Machine. *Proceeding in Neural Network for Signal Processing IX*. 23-25 August. Madison, WI, USA:IEEE, 3-11.