UNIVERSITI PUTRA MALAYSIA

A NEW CLASSIFIER BASED ON COMBINATION OF GENETIC PROGRAMMING AND SUPPORT VECTOR MACHINE IN SOLVING IMBALANCED CLASSIFICATION PROBLEM

MUHAMMAD SYAFIQ BIN MOHD POZI

FSKTM 2016 4

A NEW CLASSIFIER BASED ON COMBINATION OF GENETIC PROGRAMMING AND SUPPORT VECTOR MACHINE IN SOLVING IMBALANCED CLASSIFICATION PROBLEM

By

MUHAMMAD SYAFIQ BIN MOHD POZI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy.

February 2016

DEDICATIONS

To my lovely parents: Mohd Pozi Mohd Zaki and Rozita Abdul Aziz.
To my lovely siblings: Nur Syuhada Mohd Pozi, Nur Syarafana Mohd Pozi and Muhammad Nur Syahid Mohd Pozi

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the Degree of Doctor of Philosophy.

A NEW CLASSIFIER BASED ON COMBINATION OF GENETIC PROGRAMMING AND SUPPORT VECTOR MACHINE IN SOLVING IMBALANCED CLASSIFICATION PROBLEM

By

MUHAMMAD SYAFIQ BIN MOHD POZI

February 2016

Chairperson : Associate Professor Md Nasir Sulaiman, PhD
Faculty : Computer Science and Information Technology

In supervised learning, class imbalanced data set is a state where the class distribution is not uniform among the classes. Many classifiers fail to properly identify pattern that belongs to minority class due to most of those classifiers are built in order to minimize error rate. Hence, a biased classification model is highly anticipated as higher accuracy can always be represented by majority class.

There are two methods in dealing with imbalanced classification problem, which are based on data or algorithmic level. Data level based methods are meant to solve the imbalanced classification problem based on the idea of making both classes equal in number. However, by changing the distribution of both classes, the original classes distribution that are followed by that particular data will be violated. Algorithmic level based methods however are based on introducing new optimization task to improve the minority class classification rate, without changing the data characteristics. Nevertheless, the optimization task requires specific care in order to prevent the issue of overfitting classification model.

Therefore, a new classifier based on genetic programming (GP) and support vector machine (SVM) is proposed in this thesis in order to solve the imbalanced classification problem without changing the data properties. The idea is to use GP to optimize the SVM decision function such that the minority class classification rate is increased without sacrificing the accuracy rate for both classes. In addition, the classifier is also optimized such that it has a good generalization property. The main keys of the new classifier are based on the new kernel method, new learning metric and a new optimization algorithm in order to optimize the SVM decision function. The proposed classifier is called Support Vector Genetic Programming Machine, SVGPM.

In order to evaluate the performance of SVGPM against current methods in solving im-

balanced classification task, three experiments are conducted such as on selected standard class imbalanced benchmark data sets, intrusion detection system (IDS) data set and remote sensing data set. The SVGPM performance is compared against SVM and cost-sensitive SVM due to the superiority of SVM in dealing with imbalanced classification problem. The second experiment is by evaluating the SVGPM performance on detecting anomalous rare attacks from network intrusion data set. The SVGPM performance is compared against current methods in developing a prediction model for IDS. In the third experiment, SVGPM is evaluated on wilt disease data set from remote sensing study, to identify wilt diseased trees in high-resolution image. The SVGPM performance is compared against the previously proposed methods in mapping the regions that are covered by wilt diseased trees in Japan.

The carried out experimentation shown that SVGPM gives a very good classification rate in classifying minority class without sacrificing the accuracy rate for both classes. This is because, in the training stage, the introduced optimization task in SVGPM ensures that each minority class example is generalized into one learning concept and both classification rate for majority and minority classes are similar.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Doktor Falsafah.

PENGKELAS BERDASARKAN KOMBINASI PENGATURCARAAN GENETIK DAN MESIN SOKONGAN VECTOR DALAM MENYELESAIKAN MASALAH KETIDAKSEIMBANGAN KLASIFIKASI

Oleh

MUHAMMAD SYAFIQ BIN MOHD POZI

Februari 2016

Pengerusi : Profesor Madya Md Nasir Sulaiman, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Dalam konteks pembelajaran diselia, ketidakseimbangan kelas data adalah suatu keadaan di mana taburan kelas tidak seragam di dalam data. Oleh itu, banyak pengkelas gagal untuk mengenali corak yang berasal daripada kelas minoriti dengan tepat kerana kebanyakan pengkelas dibina untuk mengurangkan kadar kesilapan dalam mengelas sesuatu data. Oleh itu, pengkelas yang berat sebelah amatlah dijangka disebabkan ketepatan yang tinggi boleh hanya diwakili oleh kelas majoriti.

Terdapat dua kaedah dalam menyelesaikan masalah klasifikasi yang tidak seimbang, sama ada berdasarkan tahap data atau tahap algoritma. Kaedah berasaskan tahap data adalah untuk menyelesaikan masalah klasifikasi yang tidak seimbang berdasarkan idea membuat jumlah data untuk kedua-dua kelas sama. Walaubagaimanapun, menukar taburan untuk kedua-dua kelas, taburan asal kelas yang diikuti oleh data tersebut akan tercemar. Kaedah berasaskan tahap algoritma pula adalah berdasarkan dengan pengenalan tugas pengoptimum yang baru untuk meningkatkan kadar klasifikasi ke atas kelas minoriti tanpa mengubah sifat data tersebut. Walaupun begitu, tugas pengoptimum yang baru perlu dibuat secara berhati-hati untuk mengelakkan masalah model klasifikasi yang terlebih pemadanan.

Oleh itu, satu pengkelas berasaskan pengatucaraan genetik (GP) dan sokongan mesin vektor (SVM) telah diusulkan di dalam tesis ini bagi menyelesaikan masalah klasifikasi yang tidak seimbang. Ideanya ialah untuk menggunakan GP bagi mengoptimumkan fungsi keputusan SVM di mana kadar klasifikasi kelas minoriti meningkat tanpa mengorbankan kadar ketepatan bagi kedua-dua kelas. Tambahan lagi, pengkelas tersebut juga dioptimumkan bagi membuatkan ia mempunyai sifat generalisasi yang bagus. Kunci utama bagi pengkelas ini ialah berdasarkan kaedah kernel yang baru, ukuran pembelajaran yang baru dan algoritma optimum yang baru, bertujuan untuk mengoptimumkan funsi keputusan SVM. Pengkelas tersebut dipanggil sebagai sokongan pengaturcaraan

genetik mesin vektor. Secara ringkasnya, pengkelas terbaru, dikenali sebagai SVGPM.

Bagi menilai keupayaan SVGPM dengan kaedah-kaedah terkini dalam menyelesaikan masalah klasifikasi yang tidak seimbang, tiga ujikaji telah dijalankan. Eksperimen pertama berdasarkan set-set data yang digunakan sebagai penanda aras dalam menentukan keupayaan pengkelas. Ujikaji yang kedua adalah untuk menganalisis keupayaan setiap pengkelas dalam membuat sistem pengesanan pencerobohan dalam talian. Ujikaji yang ketiga pula adalah untuk menganalisis keupayaan setiap pengkelas dalam memetakan kawasan hutan yang mempunyai setiap pokok yang berpenyakit di kawasan pergunungan Jepun.

Ujikaji-ujikaji yang telah dijalankan menunjukkan keupayaan SVGPM yang sangat baik dalam mengklasifikasikan kelas minoriti tanpa mengorbankan kadar ketepatan untuk kedua-dua kelas. Ini kerana, dalam peringkat latihan SVGPM, tugas pengoptimuman dalam SVGPM memastikan bahawa setiap contoh kelas minoriti adalah umum kepada satu konsep pembelajaran dan kedua-dua kadar klasifikasi untuk kelas majoriti dan kelas minoriti adalah sama.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Assoc. Prof. Md Nasir Sulaiman for the continuous support of my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study. Besides my advisor, I would like to thank the rest of my thesis committee: Assoc. Prof. Norwati Mustapha, and Dr. Thinagaran Perumal, for their encouragement, insightful comments, and hard questions. I thank all my fellow labmates in Intelligent Computing Group. In particular, I am grateful to Dr. Hanizan Shaker for enlightening me the first glance of research. Last but not the least, I would like to thank my family: my parents Mohd Pozi Mohd Zaki and Rozita Abdul Aziz, for giving birth to me at the first place and supporting me for everything throughout my life. I am grateful for my great siblings: Nur Syuhada Mohd Pozi, Nur Syarafana Mohd Pozi and Muhammad Nur Syahid Mohd Pozi.

I certify that a Thesis Examination Committee has met on 11 February 2016 to conduct the final examination of Muhammad Syafiq bin Mohd Pozi on his thesis entitled "A New Classifier Based on Combination of Genetic Programming and Support Vector Machine in Solving Imbalanced Classification Problem" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Ali bin Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Razali bin Yaakob, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Aida binti Mustapha, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Nikola Kirilov Kasabov, PhD
Professor
Auckland University of Technology
New Zealand
(External Examiner)

ZULKARNAIN ZAINAL, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 25 May 2016

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

Md Nasir Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairperson)

Norwati Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Thinagaran Perumal, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:____________________ Date:________________

Name and Matric No.: Muhammad Syafiq Bin Mohd Pozi GS33773

Declaration by Members of Supervisory Committee

This is to confirm that:
- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________
Name of Chairman of Supervisory Committee: Associate Professor Md Nasir Sulaiman

Signature: ______________________
Name of Member of Supervisory Committee: Associate Professor Norwati Mustapha

Signature: ______________________
Name of Member of Supervisory Committee: Dr Thinagaran Perumal

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
GP	Genetic Programming
DT	Decision Tree
SVGPM	Support Vector Genetic Programming Machine
GSVM	Geometric Mean Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Introduction

Machine learning is aimed at developing a system that learn. The theoretical foundation of machine learning and its application in real world domains have been immensely explored in the last decades (Gonzalez-Abril et al., 2014; Chatrath et al., 2014; Kotsiantis, 2013; Loh, 2011; Gu et al., 2014; Chen et al., 2014).

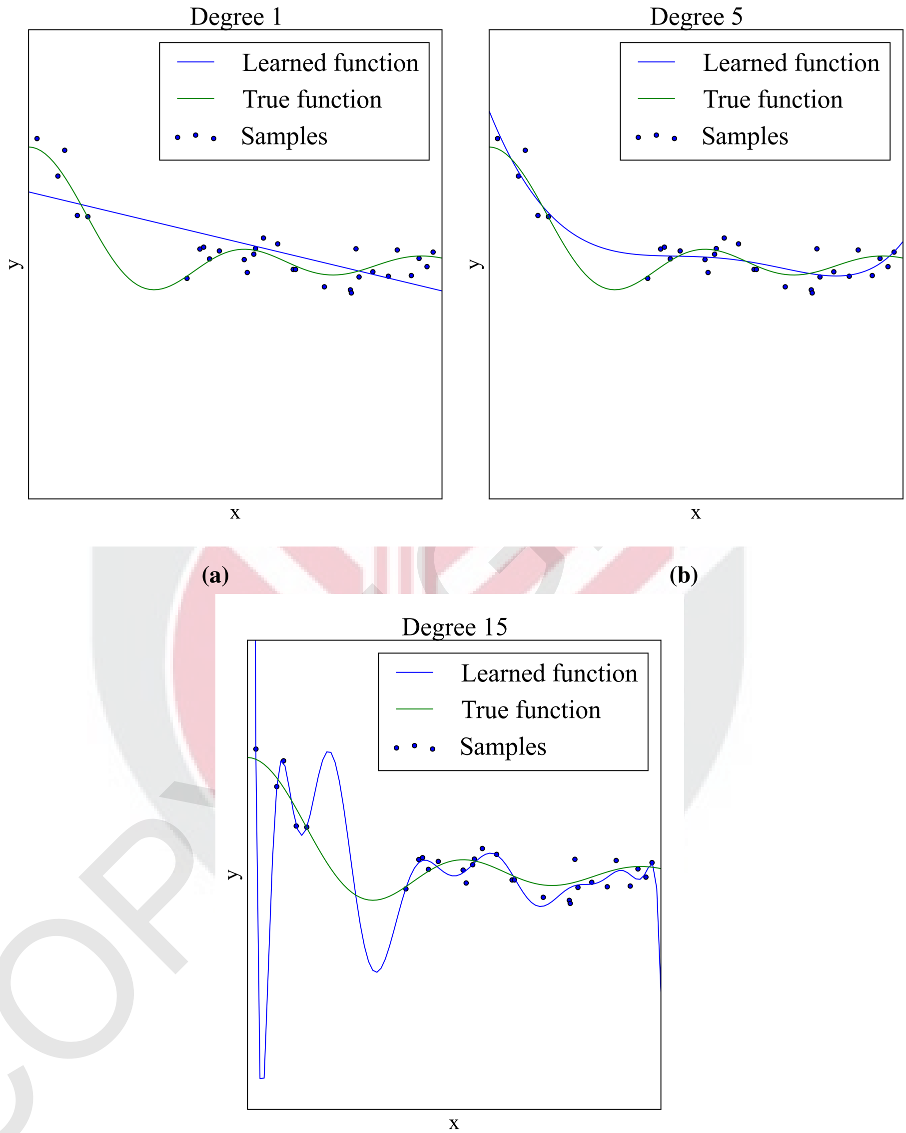
In general, machine learning tasks can be classified into two basic categories, which are supervised learning (Garcia et al., 2013; Jordan and Jacobs, 2014) and unsupervised learning (Mirkin, 2012; Karaboga and Ozturk, 2011; Zimek et al., 2014, 2013). However, the focus of this thesis is solely on supervised learning.

1.2 Supervised Learning

Supervised learning is a type of learning where an input-output relation is learned from input-output samples, which is also known as training samples. Then, the learned relation will be validated on unseen input-output samples, which also known as validation or testing samples. Two common tasks in supervised learning are classification and regression. The objective of classification task is to determine the correct discrete output given a vector of input while the objective of regression task is to predict the correct continuous value as the output given a vector of input.

Formulating a good learning algorithm for those tasks is the main research focus in supervised learning. Beyond that, there are several recent research issues in supervised learning that are quite challenging to solve such as model selection, active learning and dimensionality reduction.

Model selection revolve around on controlling the complexity of the learned function induced by learning algorithm to obtain good prediction performance. Here, a model is a set of functions where the best performing learning function is learned from, whereby the complexity of the learning function is directly related with the number of variables utilized by the function. Either way, a good learning function should generalize well on unseen input-output samples. However, this cannot be achieved if the complexity of the learned function is not controlled properly. A learning function with high complexity will result in high variance in prediction performance while a learning function with low complexity will result in high bias in prediction performance. Figure 1.1 illustrates three learning functions with different degree of complexities applied on the training samples.



(a) (b)

(c)

Figure 1.1: The effect of various degree of complexity of the learning function in approximating the true function.

Figure 1.1a, with degree of 1, the learning function is not sufficient to fit the samples. Figure 1.1b, with degree of 5, the learning function almost fit all the samples while Figure 1.1c, with degree of 15, the learning function unnecessary learns the samples noise, resulting the learning function significantly deviated from the sample true function. Hence, from Figure 1.1 it can be concluded that the increased complexity of learning function does not guarantee the function is closely related to the true function even though it fits all the training samples. Moreover, blindly increasing the complexity of the learning function will only increase the training time by large margin without any real benefit. This is why model selection is an important factor that need to be properly taken care of when formulating a new learning algorithm.

However, most of real world problems are not as simple as it seems in Figure 1.1. The complexity in finding the true function can be affected by two main factors which are the nature of the samples that need to be modelled and the formulation of the learning algorithm. This is because, there are two basic principles in supervised learning (Moreno-Torres et al., 2012; He and Garcia, 2009; Mitchell, 2009) which are:

1. **Assumption 1**: Both training and future samples have similar distribution and characteristic. In order to perform any supervised learning task, it is assumed that that the distribution of training and future samples is stationary, that is, the future samples will not change over time. Hence, the task at hand is simply to estimate the distribution of training samples. However, almost all the time, this assumption is rarely fulfilled, for example, when the area outside of the training region is extrapolated because of the nature of the data producer has been changed significantly.

2. **Assumption 2**: Common objective function for learning algorithms is to minimize error rate. Most standard learning algorithms are build based on empirical risk minimization, such that, the main objective function of learning algorithms is to minimize the error rate of the learning function. However, the learning function does not have the ability to differentiate between important input points and useless or less important input points such as noises or outliers. Depending on the samples, this important input points might be represented by lower number of points compare to useless input points. As a result, the resulting learning function performance is largely contributed by useless input points, which result in a biased learning function. In classification task, this problem is closely associated with imbalanced classification problem. Hence, imbalanced classification which is the main focus of this thesis.

Hence, several frameworks have been designed to help researcher to determine the expected performance of each learning function inside the specified learning model when one or both assumptions mentioned before are violated, especially in high dimensional samples, among other, such as Cross Validation (CV) (Arlot et al., 2010), Aikake Information Criterion (AIC) (Hu, 2007), and Structural Risk Minimize Vapnik-Chervonnenkis dimension (SRMVC) (Buhmann and Gronskiy, 2013). CV requires each learning function inside the model to be trained by the same settings. Then the learning functions are validated on the testing samples in order to determine the performance of the learned function, which is basically to minimize the empirical error rate. On the contrary, AIC and SRMVC only need the training samples which select the best learning function based on certain criteria. AIC values requires the learning algorithm to define the maximum likelihood function and the number of free parameters. The free parameters will be penalized based on the increasing function, for each learning function in the model. The best learning function is the one that has the lowest AIC value.

SRMVC however prioritize the simplest learning function over the training error. In contrast with CV and AIC, SRMVC requires some principles (Zhang, 2010) to be adhered such as follows:

1. Based on the domain prior information, pick a class of capacities, for example, polynomials of degree *n*, neural systems having *n* hidden layer neurons, an arrangement of splines with *n* hubs or fuzzy logic models having *n* rules.

2. Partition the functions into a progression of settled subsets in place of expanding multifaceted nature, for example, polynomials of increasing degree.

3. The empirical risk minimization is performed on each subset.

4. Finally the function in the series whose sum of empirical risk and Vapnik- Chervonenkis confidence is minimal, is selected as the learning function.

Nonetheless, it is almost an infeasible process to evaluate each subset in order to comply with SRMVC principles, especially when some learning functions are based on very complex learning algorithm which requires high processing computational power to make it as a feasible process. Therefore, in order to overcome this problem, a SRMVC based learning algorithm that adhere to the SRMVC principles has been developed which is called as Support Vector Machine (SVM) and its regression type which is known as Support Vector Regression (SVR).

Other research issues in formulating a learning algorithm are related to active learning and dimensionality reduction. Active learning is a type of learning where users are allow to design the location of training input points in order to maximize the performance of the learning function while dimensionality reduction is a process to reduce the complexity of input-output samples under assumption that some points in the samples are redundant or useless due to noises or outliers. Both of these issues are beyond the scope of this thesis.

1.3 Imbalanced Classification Problem

The issue of imbalanced classification problem appears often on data mining applications due to many reasons such as direct result of the nature of the dataspace, time and storage issue (He and Garcia, 2009) which make learning the distinction between classes, i.e., the true function or concept for each class, difficult.

One of the domain that always dealing with imbalanced classification problem is when modelling medical problem. For example, Gil et al. (2012b) has develop a learning model for seminal quality based on life factor. However, the accuracy of the model cannot properly discriminate bad sperm quality due to insufficient bad sperm data. As a result, even though the classification accuracy is high, but the model is bias since high accuracy can solely be represented by majority class or good sperm data.

1.4 Problem Statement

Several learning algorithms have been proposed to solve imbalanced classification problem. However, recent review on many proposed learning algorithms such as C4.5 (Quinlan, 2014), Naive-Bayes (Jiang et al., 2014; Zhang, 2004), and Neural Network (Maren et al., 2014) with respect to the mentioned learning strategies seems to suggest that they are susceptible to class imbalance. This is because, it is hard to control the generalization property of those classifiers. Both C4.5 and Neural Network are very easy to overfit, while Naive-Bayes requires the user to specify the best attributes as it can't learn the relation among the data attributes.

Several works based on SVM (Gonzalez-Abril et al., 2014; Maratea et al., 2014; Imam et al., 2006) have shown that SVM is the classifier paradigm that is less affected by class imbalance, being almost insensitive to all but the most imbalanced distributions (Prati et al., 2014). This is because, based on the SRMVC (Zhang, 2010) principles, SVM learning function has a strong generalization property as it can be represented by a smaller subset of patterns, hence, making SVM a usually preferred classifier when dealing with imbalanced classification problem.

However, the experimental results obtained from those SVM based techniques shown that there is always a compromise between the total accuracy and precision of minority class. In addition, when Assumption 1 is violated, the performance of those techniques are significantly reduced in term of specificity value on minority class. This is probably because the proposed techniques (Gonzalez-Abril et al., 2014; Imam et al., 2006) reduces the generalization property of SVM as the tradeoff that need to be paid to improve the classification rate of minority class based on standard learning metrics such as specificity or geometric mean. In addition, other proposals, such as designing a new SVM kernel (Maratea et al., 2014; Zhang et al., 2014b) requires longer computing time in the training stage due to the complexity introduced in tuning the appropriate parameters, which is introduced in the new kernel.

Thus, we are attempting to improve the classification performance on class imbalanced data set, based on SVM, without paying a significant tradeoff between accuracy and precision of the minority class without increasing the complexity in model selection, and also to improve the generalization performance on unseen data that have different distribution with training data. In this thesis, we refer the issue of data having different classes distribution between training and future data as dynamic data.

1.5 Research Objective

The primary objective of this research is to propose a new classifier in order to improve the classification rate on minority class without sacrificing the overall accuracy. In addition, the classification model from the proposed classifier can be generalized on unseen data that have different classes distribution between training and future data. In order to achieve the primary objectives, the following objectives are adopted:

1. To propose a new SVM kernel method that transform input data into higher dimensional space to solve imbalanced classification problem. The SVM is chosen due to its high generalization property, as previously mentioned in Section 1.1.

2. To formulate a new learning metric that need to be maximized in order to control the classifier complexity.

3. To propose a new evolutionary optimization algorithm based on genetic programming that use both of the proposed SVM kernel and formulated learning metric in order to improve the precision of minority class without significantly sacrificing the accuracy of the learning function and in addition to improve the generalization performance on unseen data that have different classes distribution between training and future data.

4. To show the applicability of the proposed classifier on real world application such as intrusion detection system (IDS) and remote sensing researches.

1.6 Research Scope

The scope of this work is centered around binary classification problems with a static training and validation samples. By static we mean they are fully known at the same time, unlike time series problems where data measurements are made available step by step.

In addition, negative and majority class are used interchangeably in this thesis, which representing a class with highest number of instances in a given data set, while positive and minority class are used interchangeably in this thesis, which representing a class with lowest number of instances in a given data set.

1.7 Research Contribution

Hence, the overall contribution of this research is to develop a new classifier based on SVM to solve imbalanced classification problem. The contribution can be divided into three main contributions such as follows:

1. A new kernel method for solving imbalanced classification problem.

2. A new learning metric that combines SVM internal structure in order to control the SVM complexity while also improving the classification performance on minority class.

3. A new evolutionary optimization algorithm for SVM in handling imbalanced classification problem which consists of the proposed kernel method and learning metric. We present a new algorithm to optimize the SVM learning function in solving imbalanced classification problem without significant reduction on the classification accuracy and generalization property using genetic programming. The new learning algorithm is called Support Vector based on Genetic Programming Machine (SVGPM).

1.8 Thesis Organization

In particular, Chapter 2 reviews the background study of imbalanced classification problem and different kinds of strategies that have been proposed recently in solving imbalanced classification problem. Some examples of imbalanced classification problem are also presented in the chapter. Next, Chapter 3 explains the research methodologies that are followed in this thesis in detail, such as the flow of the research, the detail of each data set that is used in the experiment and also software and hardware requirements. Then, Chapter 4 describes the new optimization algorithm based on SVM to improve the learning performance on imbalanced classification problem, resulting to a new classifier called Support Vector Genetic Programming Machine (SVGPM). Furthermore, Chapter 4 is the main contribution of the thesis. Next, Chapter 5 discusses the experimentation design, result and analysis of the obtained result of the proposed classifier. Finally, Chapter 6 concludes the thesis and suggested several improvement that can be done based on this research contribution as future work.

REFERENCES

Afzal, Z., Schuemie, M. J., van Blijderveen, J. C., Sen, E. F., Sturkenboom, M. C. and Kors, J. A. 2013. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC medical informatics and decision making* 13 (1): 30.

Aho, K., Derryberry, D. and Peterson, T. 2014. Model selection for ecologists: the world-views of AIC and BIC. *Ecology* 95 (3): 631–636.

Alpaydin, E. 2014. *Introduction to machine learning*. MIT press.

Amin, A., Rahim, F., Ali, I., Khan, C. and Anwar, S. 2015, In New Contributions in Information Systems and Technologies, In *New Contributions in Information Systems and Technologies*, 215–225, Springer, 215–225.

Arlot, S., Celisse, A. et al. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys* 4: 40–79.

Ashlock, D., Smucker, M. and Walker, J. 1999. Graph based genetic algorithms. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*. IEEE.

Bache, K. and Lichman, M. 2013, UCI Machine Learning Repository.

Barua, S., Islam, M. M., Yao, X. and Murase, K. 2014. MWMOTE–Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Trans. on Knowl. and Data Eng.* 26 (2): 405–425.

Batuwita, R. and Palade, V. 2010. FSVM-CIL: fuzzy support vector machines for class imbalance learning. *Fuzzy Systems, IEEE Transactions on* 18 (3): 558–571.

Bielza, C. and Larranaga, P. 2014. Discrete Bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)* 47 (1): 5.

Bishop, C. M. et al. 2006. *Pattern recognition and machine learning*. , vol. 1. springer New York.

Buhmann, J. M. and Gronskiy, A. 2013. Statistical Learning Theory .

Chandola, V., Banerjee, A. and Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41 (3): 15.

Chang, C.-C. and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2 (3): 27.

Chatrath, J., Gupta, P., Ahuja, P., Goel, A. and Arora, S. M. 2014. Real time human face detection and tracking. In *Signal Processing and Integrated Networks (SPIN), 2014 International Conference on*, 705–710. IEEE.

Chauhan, H., Kumar, V., Pundir, S. and Pilli, E. 2013. A Comparative Study of Classification Techniques for Intrusion Detection. In *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*, 40–43.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16: 321–357.

Chen, Y., Jalali, A., Sanghavi, S. and Xu, H. 2014. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research* 15 (1): 2213–2238.

Cristianini, N. and Shawe-Taylor, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cunningham, P. and Delany, S. J. 2007. k-Nearest neighbour classifiers. *Multiple Classifier Systems* 1–17.

Datta Chaudhuri, T., Ghosh, I. and Eram, S. 2015. Predicting Stock Returns of Mid Cap Firms in India–An Application of Random Forest and Dynamic Evolving Neural Fuzzy Inference System. *Available at SSRN 2709913* .

Demšar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7: 1–30.

Denil, M. and Trappenberg, T. 2010, In Advances in Artificial Intelligence, In *Advances in Artificial Intelligence*, 220–231, Springer, 220–231.

Deshmukh, D. H., Ghorpade, T. and Padiya, P. 2014. Intrusion detection system by improved preprocessing methods and Naïve Bayes classifier using NSL-KDD 99 Dataset. In *Electronics and Communication Systems (ICECS), 2014 International Conference on*, 1–7. IEEE.

Du, W., Han, Y. S. and Chen, S. 2004. Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification. In *SDM*, 222–233. SIAM.

Duman, E. and Ozcelik, M. H. 2011. Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications* 38 (10): 13057 – 13063.

Eggermonta, J. 2007. Genetic programming. *Leiden University Medical Center* .

Enache, A.-C. and Patriciu, V. 2014. Intrusions detection based on Support Vector Machine optimized with swarm intelligence. In *Applied Computational Intelligence and Informatics (SACI), 2014 IEEE 9th International Symposium on*, 153–158.

Fawcett, T. and Flach, P. A. 2005. A response to Webb and Tings on the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning* 58 (1): 33–38.

Fingas, M. and Brown, C. 2014. Review of oil spill remote sensing. *Marine pollution bulletin* 83 (1): 9–23.

Frank, E. and Witten, I. H. 1998. Generating Accurate Rule Sets Without Global Optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 144–151. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Friedman, J. H. 1991. Multivariate adaptive regression splines. *The annals of statistics* 1–67.

Friedman, J. H. 1997. On bias, variance, 0/1loss, and the curse-of-dimensionality. *Data mining and knowledge discovery* 1 (1): 55–77.

Fuentes, J., Romero, C., García-Martínez, C. and Ventura, S. 2014. Accepting or Rejecting Students Self-grading in their Final Marks by using Data Mining. In *Proceedings of the 7th International Conference on Educational Data Mining, EDM*.

Gagné, C. and Parizeau, M. 2006. Genericity in Evolutionary Computation Software Tools: Principles and Case Study. *International Journal on Artificial Intelligence Tools* 15 (2): 173–194.

Garcia, S., Luengo, J., Sáez, J. A., López, V. and Herrera, F. 2013. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on* 25 (4): 734–750.

Gepp, A. and Kumar, K. 2015. Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science* 54: 396–404.

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J. and Johnsson, M. 2012a. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications* 39 (16): 12564–12573.

Gil, D., Girela, J. L., Juan, J. D., Gomez-Torres, M. J. and Johnsson, M. 2012b. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications* 39 (16): 12564 – 12573.

Gonçalves, I. and Silva, S. 2013. *Balancing learning and overfitting in genetic programming with interleaved sampling of training data*. Springer.

Gonzalez-Abril, L., Nuñez, H., Angulo, C. and Velasco, F. 2014. GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems. *Applied Soft Computing* 17: 23–31.

Griggs, D. J. and Noguer, M. 2002. Climate change 2001: The scientific basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. *Weather* 57 (8): 267–269.

Gu, X., Ni, T. and Wang, H. 2014. New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification. *The Scientific World Journal* 2014.

Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A. 2008. *Feature extraction: foundations and applications*. , vol. 207. Springer.

Hao, M., Wang, Y. and Bryant, S. H. 2014. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica chimica acta* 806: 117–127.

He, H. and Garcia, E. A. 2009. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on* 21 (9): 1263–1284.

He, J., Ding, L., Jiang, L. and Ma, L. 2014. Kernel ridge regression classification. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, 2263–2267. IEEE.

Hofmann, M. 2006. Support Vector MachinesKernels and the Kernel Trick. *Hauptseminar report* 32.

Hofmann, T., Schölkopf, B. and Smola, A. J. 2008. Kernel methods in machine learning. *The annals of statistics* 1171–1220.

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. 2000. *Introduction to the logistic regression model*. Wiley Online Library.

Hou, Y. and Zheng, X. F. 2011. SVM Based MLP Neural Network Algorithm and Application in Intrusion Detection. In *Proceedings of the Third International Conference on Artificial Intelligence and Computational Intelligence - Volume Part III*, 340–345. Berlin, Heidelberg: Springer-Verlag.

Hu, S. 2007. Akaike information criterion. *Center for Research in Scientific Computation* .

Huang, J. and Liu, J. 2012. Intrusion detection system based on improved BP Neural Network and Decision Tree. In *Advanced Computational Intelligence (ICACI), 2012 IEEE Fifth International Conference on*, 188–190.

Hulten, G., Spencer, L. and Domingos, P. 2001. Mining time-changing data streams. In *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 97–106. ACM Press.

Igel, C., Heidrich-Meisner, V. and Glasmachers, T. 2008. Shark. *JMLR* 9: 993–996.

Imam, T., Ting, K. M. and Kamruzzaman, J. 2006, In AI 2006: Advances in Artificial Intelligence, In *AI 2006: Advances in Artificial Intelligence*, 264–273, Springer, 264–273.

Izenman, A. J. 2013, In Modern Multivariate Statistical Techniques, In *Modern Multivariate Statistical Techniques*, 237–280, Springer, 237–280.

Jalali, V. and Leake, D. 2013, In Modeling and using context, In *Modeling and using context*, 101–114, Springer, 101–114.

Jeya, P. G., Ravichandran, M. and Ravichandran, C. S. 2012. Article: Efficient Classifier for R2L and U2R Attacks. *International Journal of Computer Applications* 45 (21): 29–32.

Jiang, L., Li, C. and Wang, S. 2014. Cost-sensitive Bayesian network classifiers. *Pattern Recognition Letters* 45 (0): 211 – 216.

Jo, T. and Japkowicz, N. 2004. Class Imbalances Versus Small Disjuncts. *SIGKDD Explor. Newsl.* 6 (1): 40–49.

Johnson, B. A., Tateishi, R. and Hoan, N. T. 2013. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing* 34 (20): 6969–6982.

Jordan, M. I. and Jacobs, R. 2014. Supervised learning and divide-and-conquer: A statistical approach. In *Proceedings of the Tenth International Conference on Machine Learning*, 159–166.

Kantschik, W. and Banzhaf, W. 2002, In Genetic Programming, In *Genetic Programming*, 83–92, Springer, 83–92.

Karaboga, D. and Ozturk, C. 2011. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied soft computing* 11 (1): 652–657.

Kim, M.-J., Kang, D.-K. and Kim, H. B. 2015. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications* 42 (3): 1074–1082.

Kotsiantis, S. B. 2013. Decision Trees: A Recent Overview. *Artif. Intell. Rev.* 39 (4): 261–283.

Koza, J. R. 1992. *Genetic programming: on the programming of computers by means of natural selection.* , vol. 1. MIT press.

Kretowski, M. and Grześ, M. 2007. *Adaptive and Natural Computing Algorithms: 8th International Conference, ICANNGA 2007, Warsaw, Poland, April 11-14, 2007, Proceedings, Part I*, Ch. Evolutionary Induction of Decision Trees for Misclassification Cost Minimization, 1–10. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kuang, F., Xu, W. and Zhang, S. 2014. A novel hybrid KPCA and SVM with GA model for intrusion detection. *Applied Soft Computing* 18: 178–184.

Kubono, T. and Ito, S.-i. 2002. Raffaelea quercivora sp. nov. associated with mass mortality of Japanese oak, and the ambrosia beetle (Platypus quercivorus). *Mycoscience* 43 (3): 0255–0260.

Kumar, K. M., Srinivas, P. and Rao, C. R. 2012. Sequential pattern mining with multiple minimum supports by MS-SPADE. *International Journal of Computer Sciences* 9 (5).

Laurikkala, J. 2001. *Improving identification of difficult small classes by balancing class distribution*. Springer.

Li, J., Li, X. and Yao, X. 2005. Cost-sensitive classification with genetic programming. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, 2114–2121. IEEE.

Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X. and Dai, K. 2012. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications* 39 (1): 424–430.

Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. *R news* 2 (3): 18–22.

Lichman, M. 2013, UCI Machine Learning Repository.

Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning* 40 (3): 203–228.

Liu, X. 2009. A benefit-cost based method for cost-sensitive decision trees. In *Intelligent Systems, 2009. GCIS'09. WRI Global Congress on*, 463–467. IEEE.

Liu, X.-Y., Wu, J. and Zhou, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39 (2): 539–550.

Liu, Y. and Khoshgoftaar, T. 2004. Reducing overfitting in genetic programming models for software quality classification. In *null*, 56–65. IEEE.

Loh, W.-Y. 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (1): 14–23.

López, V., Fernández, A. and Herrera, F. 2014. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences* 257: 1–13.

Mani, I. and Zhang, I. 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Maratea, A., Petrosino, A. and Manzo, M. 2014. Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences* 257: 331–341.

Maren, A. J., Harston, C. T. and Pap, R. M. 2014. *Handbook of neural computing applications*. Academic Press.

Márquez-Vera, C., Cano, A., Romero, C. and Ventura, S. 2013. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence* 38 (3): 315–330.

McHugh, J. 2000. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM transactions on Information and system Security* 3 (4): 262–294.

Mease, D., Wyner, A. J. and Buja, A. 2007. Boosted classification trees and class probability/quantile estimation. *The Journal of Machine Learning Research* 8: 409–439.

Mirkin, B. 2012. *Clustering: a data recovery approach*. CRC Press.

Mitchell, T. 2009. Machine Learning. *ISBN: 0-07-042807-7, Publisher: McGraw Hill* .

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45 (1): 521–530.

Morrison, D. 2007. Nbtree: A naive bayes/decision-tree hybrid .

Nakamura, Y., Ghaibeh, A. A., Setoguchi, Y., Mitani, K., Abe, Y., Hashimoto, I. and Moriguchi, H. 2015. On-Admission Pressure Ulcer Prediction Using the Nursing Needs Score. *JMIR medical informatics* 3 (1).

Ohta, K., Hoshizaki, K., Nakamura, K., Nagaki, A., Ozawa, Y., Nikkeshi, A., Makita, A., Kobayashi, K. and Nakakita, O. 2012. Seasonal variations in the incidence of pine wilt and infestation by its vector, Monochamus alternatus, near the northern limit of the disease in Japan. *Journal of forest research* 17 (4): 360–368.

Orriols-Puig, A., Bernadó-Mansilla, E., Goldberg, D. E., Sastry, K. and Lanzi, P. L. 2009. Facetwise analysis of XCS for problems with class imbalances. *Evolutionary Computation, IEEE Transactions on* 13 (5): 1093–1119.

Park, E.-m. and Lee, J.-H. 2013. Classifying imbalanced data using an Svm ensemble with k-means clustering in semiconductor test process. In *Sixth International Conference on Machine Vision (ICMV 13)*, 90672D–90672D. International Society for Optics and Photonics.

Peterson, L. E. 2009. K-nearest neighbor. *Scholarpedia* 4 (2): 1883.

Ponz, A., Rodríguez-Garavito, C., García, F., Lenz, P., Stiller, C. and Armingol, J. 2015. Laser Scanner and Camera Fusion for Automatic Obstacle Classification in ADAS Application. In *Smart Cities, Green Technologies, and Intelligent Transport Systems: 4th International Conference, SMARTGREENS 2015, and 1st International Conference VEHITS 2015, Lisbon, Portugal, May 20-22, 2015, Revised Selected Papers*, 237–249. Springer.

Prati, R. C., Batista, G. E. and Monard, M. C. 2004, In MICAI 2004: Advances in Artificial Intelligence, In *MICAI 2004: Advances in Artificial Intelligence*, 312–321, Springer, 312–321.

Prati, R. C., Batista, G. E. and Silva, D. F. 2014. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems* 1–24.

Puertas, O. L., Brenning, A. and Meza, F. J. 2013. Balancing misclassification errors of land cover classification maps using support vector machines and Landsat imagery in the Maipo river basin (Central Chile, 1975–2010). *Remote Sensing of Environment* 137: 112–123.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1 (1): 81–106.

Quinlan, J. R. 2014. *C4. 5: programs for machine learning*. Elsevier.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. 2009. *Dataset shift in machine learning*. The MIT Press.

Ren, D., Yu, H., Fu, W., Zhang, B. and Ji, Q. 2012. Crop diseases and pests monitoring based on remote sensing: A survey. In *World Automation Congress (WAC), 2012*, 177–181. IEEE.

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M. and Verri, A. 2004. Are loss functions all the same? *Neural Computation* 16 (5): 1063–1076.

Sahin, Y., Bulkan, S. and Duman, E. 2013. A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications* 40 (15): 5916 – 5923.

Salvatore J. Stolfo, Wei Fan, W. L. A. P. and Chan, P. K. 1999. *KDD Cup 1999 Data*.

Shi, H. 2007, Best-first decision tree learning, Tech. rep., University of Waikato.

Srinivas, K., Rao, G. R. and Govardhan, A. 2012. Mining association rules from large datasets towards disease prediction. In *2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT*.

Tavallaee, M., Bagheri, E., Lu, W. and Ghorbani, A.-A. 2009. A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*.

Tesfahun, A. and Bhaskari, D. L. 2013. Intrusion Detection Using Random Forests Classifier with SMOTE and Feature Reduction. In *Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on*, 127–132. IEEE.

Thammasiri, D., Delen, D., Meesad, P. and Kasap, N. 2014. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* 41 (2): 321–330.

Trimble. 2011, eCognition Developer 8.7: Reference Book.

Tsai, C.-F. and Lin, C.-Y. 2010. A triangle area based nearest neighbors approach to intrusion detection. *Pattern Recognition* 43 (1): 222–229.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Veropoulos, K., Campbell, C., Cristianini, N. et al. 1999. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on artificial intelligence*, 55–60.

Wan, X., Liu, J., Cheung, W. K. and Tong, T. 2014. Learning to improve medical decision making from imbalanced data without a priori cost. *BMC medical informatics and decision making* 14 (1): 111.

Wang, J., Feng, J. and Han, Z. 2014. Fault detection for the class imbalance problem in semiconductor manufacturing processes. *Journal of Circuits, Systems, and Computers* 23 (04): 1450049.

Wang, X.-Y., Yang, L., Liu, R. and Kadir, A. 2010. A chaotic image encryption algorithm based on perceptron model. *Nonlinear Dynamics* 62 (3): 615–621.

Weiss, G. M. 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter* 6 (1): 7–19.

Weiss, G. M. 2010. The impact of small disjuncts on classifier learning. In *Data Mining*, 193–226. Springer.

Weiss, G. M. and Provost, F. 2001. The effect of class distribution on classifier learning: an empirical study. *Rutgers Univ* .

Widrow, B. and Lehr, M. A. 1990. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE* 78 (9): 1415–1442.

Witten, I. H. and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, J. and Zhang, X. 2001. A PCA classifier and its application in vehicle detection. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, 600–604. IEEE.

Zhang, H. 2004. The optimality of naive Bayes. *AA* 1 (2): 3.

Zhang, X. 2010, In Encyclopedia of Machine Learning, In *Encyclopedia of Machine Learning* (eds. C. Sammut and G. Webb), 929–930, Springer US, 929–930.

Zhang, X., Song, Q., Zheng, Y., Hou, B. and Gou, S. 2014a. Classification of imbalanced hyperspectral imagery data using support vector sampling. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, 2870–2873. IEEE.

Zhang, Y. and Clark, S. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational linguistics* 37 (1): 105–151.

Zhang, Y., Fu, P., Liu, W. and Chen, G. 2014b. Imbalanced data classification based on scaling kernel-based support vector machine. *Neural Computing and Applications* 25 (3-4): 927–935.

Zheng, Z. and Webb, G. 2000. Lazy Learning of Bayesian Rules. *Machine Learning* 4 (1): 53–84.

Zhou, M., Hall, L. O., Goldgof, D. B., Gillies, R. J. and Gatenby, R. A. 2015. Imbalanced learning for clinical survival group prediction of brain tumor patients. In *SPIE Medical Imaging*, 94142K–94142K. International Society for Optics and Photonics.

Zhu, L., Zhang, Z. and Zhang, D. 2015. Recognising lepidopteran images based on locality-constrained linear coding and SVM. *Oriental Insects* 1–11.

Zimek, A., Campello, R. J. and Sander, J. 2014. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *ACM SIGKDD Explorations Newsletter* 15 (1): 11–22.

Zimek, A., Gaudet, M., Campello, R. J. and Sander, J. 2013. Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 428–436. ACM.

Zughrat, A., Mahfouf, M., Yang, Y. and Thornton, S. 2014. Support Vector Machines for Class Imbalance Rail Data Classification with Bootstrapping-based Over-Sampling and Under-Sampling. In *19th World Congress of the International Federation of Automatic Control. Cape Town*.