Imbalanced Data Classification Using SVM Based on Simulated Annealing Featuring Synthetic Data Generation and Reduction

by

Hussein Ibrahim Hussein
(1740212607)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Electronic Engineering Technology
UNIVERSITI MALAYSIA PERLIS

2021

ACKNOWLEDGEMENT

All praises and thanks are due to the Almighty Allah who always guides me to the right path and has helped me to complete this thesis. There are many people whom I have to acknowledge for their support, help and encouragement during the journey of preparing this thesis. So, I will endeavor to give them their due here.

First and foremost, I would like to express my appreciation to my supervisor Dr. **Said Amirul Anwar** for his supervision, also my co-supervisor prof. Muhammad Imran Ahmad. Advice and guidance from the early stage of my research. He has given me extraordinary experiences throughout the work. Above all and the most needed, he provided me persistent encouragement and support in various ways. I am really indebted to him more than he knows.

I extend My gratitude to dean of faculty of electronic engineering technology prof. Azremi Abdullah Al-Hadi. Many thanks also go to all staff in school of Computer and communication engineering in Universiti Malaysia Perlis.

I wish to express my thanks and gratitude to my parents, the ones who can never ever be thanked enough, for the overwhelming love and care they bestow upon me, and who have supported me financially as well as morally and without whose proper guidance it would have been impossible for me to complete my higher education.

My special, profound and affectionate thanks, love, affectionate gratitude and deep indebtedness are due to my mother, who has been struggling with me, hand by hand, to secure and shape brighter future. Her understanding, support, commitment and looking after me during my study all stand behind my success. Also, all thanks and appreciation to my late father, the economic expert (**Ibrahim Hussein Sarhan**).

At the same time, I would like to express my love and thanks to 'the beats of my heart,' my brothers, **Ammar, Karar,** and **Mohammed**, who are the only source of inspiration to me, and it is their love and honest cares that have made the hardship of this task bearable. Finally, I would like to express my sincere gratitude to all my friends in Malaysia who have helped me much and without their support and prayers I would never have been able to accomplish this task.

To my beloved mother
"I would never achieve this without you"

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support vector machine
OVO	One versus one
OVA	One versus all
OSH	Optimal separating hyperplane
KKT	Karush Kuhn tucker
SM	Soft margin
PREE	Prediction risk-based feature selection for easy ensemble
CS	Cost sensitive
US	Under sampling
AM	Algorithm modification
OS	Over sampling
OSS	One sided selection
CNN	Convolutional neural network
NCL	Neighborhood cleaning rule
ENN	Extended nearest neighbor
ELM	Ensemble learning method
MSE	Mean squared error
RF	Random forest
BRF	Balanced random forest
WRF	weighted random forest
NB	Naïve bayes
KNN	K-nearest neighbor
GA	Genetic algorithm
SMO	Sequential minimal optimization
NHCMC	Nonparallel Hyperplanes Support Vector Machine for Multi-class
MBSVM	Multiple birth support vector machine
LSSVM	Least squares support vector machine
MCSVM	Multi-class Support Vector Machine
ECOC	Error Correcting Output Codes
QQP	quadratic programming problem
TWSVM	Twin Support Vector Machine
SA	Simulated Annealing
LSTSVM	Least squares twin support vector machine
NHSVM	Nonparallel hyperplane support vector machine
NLSSVM	Normal least squares support vector machine
LSNHSVM	Least squares nonparallel hyperplanes support vector machine
MLSTSVM	Multiclass least squares twin support vector machine
GA	Generate of dataset based on synthetic sampling
DR	Data Reduction
SA	Simulated Annealing

Pengelasan Data Tidak Seimbang Menggunakan SVM Berdasarkan Penyepuhlindapan Simulasi Dengan Berasaskan Penjanaan Data Tiruan dan Pengurangan

ABSTRAK

Pengelasan adalah aspek asas pembelajaran mesin yang merujuk kepada proses mengenal pasti unsur yang berasal dari salah satu daripada banyak kategori yang ada. Biasanya, proses ini mengandaikan bahawa jumlah data semasa fasa latihan adalah sama untuk setiap kelas. Namun, ini tidak berlaku ketika berhadapan dengan masalah pengelasan dunia-nyata di mana jumlah data dalam satu kelas mungkin lebih sedikit berbanding yang lain di mana keadaan ini menyebabkan set data menjadi tidak seimbang. Isu set data yang tidak seimbang ini akan menyebabkan pengelasan mengalami masalah dengan kelas minoriti kerana bilangan sampel kelas majoriti yang lebih tinggi. Dalam tesis ini, pengelasan berasaskan SVM yang digabungkan dengan algoritma penyepuhlindapan simulasi yang melibatkan penjanaan dan pengurangan data sintetik dicadangkan untuk meningkatkan pengelasan data yang tidak seimbang. Algoritma ini dicadangkan bertujuan untuk menyelesaikan strategi pengambilan sampel bagi menangani set data yang tidak seimbang di mana pertamanya data tiruan dihasilkan untuk mengimbangkan jumlah data antara kelas dan diikuti dengan algoritma pengurangan untuk menghilangkan data lebihan dan pendua. Oleh kerana SVM memerlukan penyesuaian parameter penalti yang tepat, algoritma penyepuhlindapan simulasi digunakan untuk memilih parameter ini. Berdasarkan kerja-kerja ujikaji yang menggunakan beberapa set data piawai, algoritma yang dicadangkan dalam tesis ini memberikan ketepatan klasifikasi yang lebih tinggi berbanding dengan SVM piawai dan beberapa algoritma yang terdapat dalam kajian lain. Dengan purata ketepatan sebanyak 87.29% untuk set data kelas binari dan rata-rata 93.59% ketepatan untuk set data pelbagai kelas telah membuktikan prestasi algoritma yang telah dicadangkan.

Imbalanced Data Classification Using SVM Based on Simulated Annealing Featuring Synthetic Data Generation and Reduction

ABSTRACT

Classification is a fundamental aspect of machine learning which refers to the act of identifying elements as being from one of the many existing categories. Typically, this task assumed that the amount of data in the training phase are equal for each class. However, this is not a case when dealing with a real-world classification problem where the amount of data in one class could be less that the other which in turn leads to the imbalanced dataset. Issues with imbalanced dataset might leads to the classifier having problems with the minority class because of the relatively higher number of majority class samples. In this thesis, an SVM-based classifier combined with simulated annealing algorithm featuring synthetic data generation and reduction was proposed to improve the classification of an imbalanced data. This algorithm was proposed primarily aims at solving resampling strategy of handling imbalanced dataset where the data were first synthetically generated to equalize the amount of data between classes and follows by reduction algorithm to remove redundant and duplicated data points. As SVM requires proper tuning of its penalty parameter, the simulated annealing was employed to accurately select this parameter. Based on experimental works using several standard datasets, the proposed algorithm provides higher classification accuracy compared to standard SVM and some of algorithms found in the literature. Registering at an average of 87.29% of accuracy for binary class datasets and at an average of 93.59% of accuracy for multi-class datasets demonstrated the good performance of the proposed works.

CHAPTER 1 : INTRODUCTION

1.1 Introduction

Classification; the act of identifying elements as being from one of many existing categories; is a fundamental aspect of machine learning. The objective of classification is to build a model or classifier with which to predict the class of incoming observations based on a set of training data with given target labels (Vergara, Mayer, Damaraju, Kiehl, & Calhoun, 2017).

Classification determines a class label with the best possible accuracy (Yang et al., 2017). Classification has captured the attention of academicians and researchers over the past few decades due to its growing significance in numerous modern applications (Alazaidah, Thabtah, & Al-Radaideh, 2015). The primary objective of classification is to learn a function from a training data set and apply it to a testing set in order to evaluate the efficacy of the learned function.

A binary classification is a classification system that comprises of two classes. Disease diagnosis is an example of a binary classification problem. When using binary classification for disease diagnosis, a classifier is fed clinical patient data in order to determine if a patient has a specific disease. The two classes correspond to either the presence or absence of a disease. Therefore, if the classes represent only the absence or presence of a particular disease property, it is considered a binary classification system (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2013). However, most practical problems require more than two classes for accurate classification.

The classification process is prone to imbalanced learning. This problem has gained widespread attention; practitioners and researchers want to address this issue since there is a high likelihood that real-world data is imbalanced. Practical problems often have very few elements from the minority class and a significantly large number of elements from the other classes. It may be challenging to use typical classification techniques used for supervised learning for precise classification of the minority class. A majority of such algorithms are built on that the dataset is balanced (Mathew, Pang, Luo, & Leong, 2017).

Furthermore, the classification process is prone to imbalanced learning. This has not gone unnoticed by practitioners and researchers and has prompted them to address imbalanced learning issues as there is a high likelihood that real-world data may be imbalanced. As practical problems often have very few elements from minority classes and a significantly larger number of elements from other classes, precisely classifying minority classes using traditional classification techniques for supervised learning proves challenging. This is because a majority of these traditional algorithms are built on the assumption that the dataset is balanced (Mathew, Pang, Luo, & Leong, 2017).

Moreover, traditional classification techniques not only produce generalisations using sample data but the most hyperplane generalisations that are deemed the best fit for the data. This inherent bias is evident in several machine learning techniques, such as the k-nearest neighbours (KNN) algorithm, support vector machine (SVM) learning models, and decision tree model. Therefore, when these techniques are used on imbalanced datasets, it is highly likely that the elements of the majority class will dominate the algorithm and the minority class elements will be overlooked, thereby leading to improper classification. This is because traditional classification techniques use a biased

theory; which assumes that most of the elements are from the majority class; in order to reduce the general classification error rate (Thabtah, Hammoud, Kamalov, & Gonsalves, 2020).

Multiclass classification is also a significant issue in machine learning as it typically handles problems using two techniques: problem adaption and problem transformation. During problem transformation, the scheme modifies the problem to a binary system. In contrast, problem adaption schemes process multiclass data using specific algorithms. A classifier for a multiclass system can be produced in two ways: (i) by using all the data and assuming that it is a single optimisation problem or (ii) by splitting the multiclass problem into binary problems. Splitting a problem into many binary problems is relatively useful as it is simpler and more effective to use binary classifiers. Furthermore, there are potent algorithms, such as SVM models; with which to process binary problems (Gupta & Gupta, 2019).

SVM-based classifiers are more likely to produce accurate results than other popular algorithms, such as decision tree models, the maximum likelihood estimation (MLE) method, and techniques that are based on neural networks (Yunqiang Zhang & Zhang, 2015; S.-J. Wu, Pham, & Nguyen, 2017). This is because SVM models attempt to determine the hyperplane separating two classes by using samples that lie closest to the edge of distributions, i.e., the support vectors. The ideal hyperplane separating two classes is situated at the maximum distance from the support vectors. As such, SVM models can use such an orientation to provide highly accurate generalisations on seen elements in terms of classifiers, such as neural networks; to reduce training error. However, in order to use an SVM classification method, a few samples need to be situated at the edges of the distributions (support vectors) in the feature space to determine the

decision surface. This requirement distinguishes SVM models from other statistical classification techniques, such as the MLE method which requires the entire training sample set to characterise the class as well as a substantially larger dataset to obtain a precise classification.

Furthermore, in some cases, it may be more feasible to identify data sets that are the most appropriate for training as well as providing support vectors prior to classification (Maldonado & López, 2018; H. Yu et al., 2015). As SVM models can perform classification using much smaller data sets, researchers may benefit from substantial savings in terms of training data collection (Maldonado, Weber, & Famili, 2014). It is also noteworthy that although SVM-based classifiers are essentially binary classifiers, they can be expanded to handle situations that comprise of multiple classes. This is because multiclass problems are typically split into several binary problems that can be analysed using a one-against-one (OAO) approach or a one-against-all (OAA) approach (Silva & Villela, 2020).

In short, it is evident that better and more efficient classification techniques are required to handle class imbalance problems. Therefore, this study developed an SVM-based classification system for handling imbalanced datasets.

1.2 Problem statement

Prediction and classification are two critical aspects of machine learning. To that end, numerous techniques for classification formulation and prediction frameworks, such as (Mordelet & Vert, 2014), bagging (X.-Z. Wang et al., 2014), and cost-sensitive learning (X.-Z. Wang, Zhang, & Wang, 2017); have been developed. Datasets with

numerous sample space attributes not only require time consuming training but increase the likelihood of high feature correlation and redundancy. The presence of duplicate samples also jeopardises classification performance (Bach, Werner, Żywiec, & Pluskiewicz, 2017) and leads to overfitting (L. Yu & Liu, 2003; X.-Z. Wang, Wang, & Xu, 2017). As such, numerous studies have attempted to select appropriate parameters for an SVM model.

A class-imbalanced dataset occurs when there is a substantial difference between the number of elements in one class and the number of elements of the other classes in relation to the sample space. The minority samples hold crucial information because there are many situations in which imbalance occurs between classes, such as satellite image classification, risk management, and medical diagnosis (Hussein & Anwar, 2021). In class-imbalanced datasets that are significantly skewed; where the majority class samples tremendously outnumber the minority class samples, traditional classification methods are not expected to yield precise results due to global accuracy problems. In such cases, traditional classifiers provide poor results for minority class elements (Zięba, Świątek, & Lubicz, 2014).

There are three primary challenges when classifying imbalanced data sets. Firstly, the resampling techniques that are traditionally used to handle imbalances are unable to retain associations between features resulting in a significant gap between the original and resampled feature spaces. For instance, experiments by Maldonado, López, & Vairetti (2019) found that although the synthetic minority oversampling technique (SMOTE) attaches a higher weight to minority classes, it was not very useful in the case of imbalanced datasets.

Secondly, the feature sets selected for minority class data classification may not be appropriate if class imbalances are not taken into consideration during feature selection. To that end, Maldonado proposed a backward elimination approach that uses successive holdout steps, where the contribution of each step is measured using a balanced-loss function obtained from an independent subset (Maldonado et al., 2014). Although the feature selection scheme may be adjusted to disregard all negative features, some researchers assert that classifiers that only use positive features may potentially replicate features for new settings where the underlying class is different. Furthermore, other researchers posit that imbalanced data sets that comprise of many negative features are invaluable in terms of real-world experience. These researchers also found that negative features are integral to precise classification as feature selection performance decreased when negative features were eliminated (Yijing, Haixiang, Xiao, Yanan, & Jinling, 2016).

Thirdly, the penalty parameter of SVM models; which was first introduced by Cortes and Vapnik (1995); is a classification technique that is based on the structural risk minimisation (SRM) algorithm. Due to its success, the SRM algorithm was quickly adopted for many classification problems. However, setting parameters with the best values proves a challenge when the SRM algorithm is used in SVM models. A key parameter (*C*) is a regularisation parameter that controls the compromise between maximising the margin and minimising the number of training set errors. As such, the accuracy of an SVM classifier significantly depends on *C* as improper selection may affect the classification results (Tharwat & Gabel, 2019). As it is essential to optimise *C* in order obtain a well-tuned SVM classifier, a penalty factor is employed to regulate classification errors and generalisation capability (C. Zhang, Zhou, Guo, Wang, & Wang, 2019).

This present study proposed an approach with which to optimise an SVM model for handling imbalanced datasets. The proposed technique uses synthetic sampling to create the dataset that handles the issue of the minority class. Referred to as data reduction, the method reduces redundancy at the instance level corresponding to the imbalanced set. The simulated annealing process is then used to optimise the SVM parameters in order to provide higher classification accuracy.

1.3 Research Question

a) What effect does a sampling method have on the distribution of majority and minority classes?

b) What effect does a data reduction method have on classification performance?

c) What are the most effective strategies of increasing the accuracy of SVM models?

1.4 Objectives

This present study aimed to propose approach that dealing with imbalanced class problems, eliminate instances of data duplication as well as increase the efficiency of SVM models by selecting appropriate parameters. As such, the objectives of this research were as follows:

a) To formulate a synthetic data generation and data reduction pre-process to mitigate data redundancy and imbalanced datasets.

b) To enhance the SVM penalty parameter using the simulated annealing algorithm as well as extended the workability of the approach to multiclass classification.

c) To verify and evaluate the performance of the proposed approach.

1.5 Research Scope

Despite the efficacy of existing learning techniques for handling imbalanced datasets, some of these techniques are difficult to use. For instance, when some of these techniques are employed on complex imbalanced datasets, the algorithm is dominated by the majority class elements. It is, therefore, vital to consider the interaction and performance of imbalanced dataset learning techniques with classification techniques.

The technique proposed in this study intended to address three significant imbalanced dataset issues by using an SVM model to solve the imbalanced problem as well as eliminate data duplication then increase the efficiency of the SVM model by choosing the appropriate parameter. To that end, the linear kernel function was employed for classification as it is commonly used when dealing with a large number of features in a particular dataset. Furthermore, it is faster to train an SVM model using the linear kernel function as only the *C* regularisation parameter needs to be optimised. Other kernels, on the other hand, require the other parameter to be optimised as well resulting in longer search times. The imbalanced dataset that was used in this study was obtained from the UCI Machine Learning Repository which is widely used by researchers all over the world

as a primary source of machine learning datasets. The fact that UCI Machine Learning Repository has been cited over 1000 times in multiple journals is a testament to its popularity. The proposed framework was created using Python programming language. As it is a general-purpose, high level, open-source programming language, Python is an excellent choice for rapid application development (RAD) and increased productivity. This study also compared the efficiency and performance of the proposed technique with that of several other techniques.

1.6 Significance of the Research

The results and findings of this study are expected to benefit classification systems. As classification systems are integral to several fields, such as medicine, image recognition, bioinformatics, and text classification among others; The aim of the present study was to enhancement of accuracy and performance by formulate an approach capable of addressing imbalanced data problems as well as eliminate instances of data duplication as well as increase the efficiency of SVM models by selecting appropriate parameters.

The imbalanced dataset sampling method proposed in this study was also integral as the number of majority class elements are usually substantially higher than the number of minority class elements. These minority class samples contain valuable insight, even in imbalanced class situations. Although many enhanced classification algorithms have been developed to address this issue, they do not work effectively on imbalanced datasets. To that end, the present study proposed using synthetic sampling to create a dataset for the minority class using the features of minority class samples.

Furthermore, the data reduction technique proposed in this study used a rough set theory and lower approximation to reduce incidents of redundancy specific to every feature of the data set thereby leading to better classification. Most extant studies only use feature selection to enhance the performance of a particular classification. However, it is noteworthy that because the proposed technique requires the selection of feature sets that are appropriate for majority class classification, the classification performance of the minority class may be reduced.

The last significant aspect of this present study is the technique proposed for regulating SVM parameters using the simulated annealing process. As an SVM model works by seeking to determine a hyperplane that separates the classes, the algorithm uses training samples situated at the edges of the class distributions. The penalty parameter also profoundly affects the classification process.

1.7 Thesis Organisation

This dissertation is structured as follows: Chapter 2 comprises a literature review of class imbalance learning, rough set theory, data sampling, and the feature selection techniques required to determine subsets. Simulated annealing-based algorithms as well as a hybrid technique for choosing SVM parameters are also discussed. Chapter 3 provides an introduction while sections 3.2 and 3.3 discuss the research framework and the proposed technique, respectively. Chapter 4 examines the learning ability of the proposed technique in several class imbalance settings. The experiments that were conducted to validate the proposed technique using a UCI dataset as well as data pre-processing and how parameter selection critically influences performance are also discussed. The performance of the proposed technique was also compared with that of