Diversity Injected Ensemble Model for Small Sample-sized Classification Problems

by

OOI BOON PIN
(1740612352)

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

Faculty of Electrical Engineering & Technology
UNIVERSITI MALAYSIA PERLIS

2024

ACKNOWLEDGMENT

I would like to take this opportunity to express the profound gratitude and deep regards to my project supervisor, Assoc. Prof. Dr. Norasmadi Abdul Rahim who has guided and coordinated me at every aspect of this project especially for the enlightenment of interest in the field of machine learning and pattern recognition. I would also like to thank my co-supervisors, Dr. Maz Jamilah for her aid in understanding complex Mathematics topics, and Assoc. Prof. Dr. Ammar Zakaria for his support throughout the research.

Also, to Professor Dr. Paulraj Murugesa Pandiyan, thanks to his conscientious teaching in C programming and Artificial Intelligence courses, encouraged me in taking on this challenge that requiring in-depth programming knowledge and machine learning theory.

A special thanks to my beloved Qing YuanYuan, who supported me, encouraged me, tolerated me, loved me, and always by my side throughout this long and hard journey, always checking on my progress to ensure I am not slacking off. It was the relationship with her that kept me motivated in completing this thesis. My greatest appreciation.

At last, I would offer my warmest regards and blessing to all of those who supported me and guide me along this project. A heart fully thanks. Have a wonderful day.

This research was supported by the Ministry of Higher Education, Malaysia, through the Fundamental Research Grant Scheme (Ref: FRGS/1/2020/ICT03/UNIMAP/02/2).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADA	Adaptive Boosting ensemble
ANN	Artificial neural network
BAG	Bootstrap aggregating ensemble
CART	Classification and Regression Tree
DropELE	Dropout Extreme Learning Ensemble
EDA	Exploratory Data Analysis
EEG	Electroencephalography
ELE	Extreme Learning Ensemble
ELM	Extreme Learning Machine
HMD	Hyaline Membrane Disease
KNN	K-Nearest Neighbor
LDC	Linear Discriminant Classifier
LR	Logistic Regression
MCS	Multiple Classifier System
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NFL	No Free Lunch
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RDS	Respiratory Distress Syndrome
RF	Random Forest
RLO	Random Linear Oracle
RO	Random Oracle
RSM	Random Subspace Method
RSO	Random Subspace Oracle
SEN	Sensitivity
SPE	Specificity
SSS	Small Sample-Sized
SVM	Support Vector Machine
TREE	Decision tree
WEKA	Waikato Environment for Knowledge Analysis
WSR	Wilcoxon Signed-Ranks

LIST OF SYMBOLS

α	Level of significant
$\boldsymbol{B}$	Base classifiers in an ensemble
b	Bias
C	Number of classes
$\boldsymbol{E}$	Ensemble model
H_0	Null hypothesis
H_1	Alternative hypothesis
K	Number of regions/nearest neighbors
L	Ensemble size
M	Number of features
N	Number of instances
Q	Q statistic
R	Ranking of results
$\boldsymbol{\omega}$	Class vector
$\boldsymbol{w}$	Weight vector
$\boldsymbol{x}$	Feature vector
χ^2	Chi-square

Model Ensembel Suntikan Kepelbagaian untuk Masalah Pengelasan Bersaiz Sampel Kecil

ABSTRAK

Penyelidik kadangkala perlu menggunakan data sampel bersaiz kecil (SSS). Data SSS cenderung untuk kurang melatih algoritma pembelajaran mesin, menjadikannya tidak berguna, kerana saiz sampelnya yang sangat kecil. Nisbah cerapan-kepada-ciri yang rendah, atau bilangan ciri yang terlalu banyak berbanding dengan bilangan cerapan yang sedikit, akan menyebabkan algoritma pengelasan terlalu lampau padan dalam beberapa keadaan ekstrem berkaitan isu data SSS. Kajian ini mencadangkan dua pendekatan: *Random Subspace Oracle* (RSO) melalui penghibridan Kaedah Subruang Rawak (RSM) dengan ensembel Oracle Linear Rawak (RLO), dan pelaksanaan baharu pendekatan keciciran dengan ensembel mesin pembelajaran ekstrem (DropELE), kedua-duanya memberi tumpuan kepada penyelesaian masalah pengelasan saiz sampel kecil. Menurut ujian berperingkat-tanda Wilcoxon tak berparameter, hasil eksperimen menunjukkan bahawa ensembel RSO mempunyai prestasi lebih baik berbanding pokok keputusan tunggal dan pengelas diskriminan linear, pada aras keertian 0.05. Dalam situasi yang melibatkan pengelas tunggal lain, prestasi ensembel RSO adalah tidak optimum. Perbandingan antara model ensembel yang sudah mapan telah dijalankan dalam eksperimen yang berikutnya. Keputusan menunjukkan bahawa model RSO menunjukkan prestasi yang sebanding dengan model ensembel lain, dengan pokok keputusan sebagai pengelas asas. Algoritma DropELE dicadangkan sebagai penyelesaian mengatasi kesulitan lampau padan yang sering dihadapi dalam masalah pengelasan data SSS dengan mengurangkan kerumitan setiap pengelas asas dalam ensembel. Kajian ini menunjukkan bahawa nisbah keciciran yang lebih tinggi dan nisbah cerapan-kepada-ciri yang ditakrifkan dengan baik dapat meningkatkan prestasi algoritma yang dicadangkan. Berdasarkan data eksperimen, ketepatan ujian meningkat secara signifikan daripada 88.9% kepada 98% apabila nisbah keciciran dinaikkan daripada 0.1 kepada 0.9. Adalah penting bahawa bilangan neuron tersembunyi dan saiz ensembel dapat memberikan impak yang signifikan. Dalam kajian ini, algoritma DropELE menunjukkan peningkatan bilangan neuron tersembunyi yang mencapai suatu ambang memberi kesan positif terhadap prestasi pengelasan. Namun, melampaui suatu ambang, ketepatan pengelasan akan mulai menurun. Pemerhatian serupa dapat dibuat berkaitan dengan saiz ensembel. Prestasi dapat ditingkatkan dengan menambah bilangan pengelas asas dalam takungan ensembel. Adalah penting untuk diakui bahawa peningkatan ini akan berkurangan secara beransur-ansur seiring dengan penambahan lebih banyak pengelas asas ke dalam takungan. Empat set data perubatan dunia sebenar telah digunakan untuk menilai prestasi dua algoritma yang dicadangkan. Kaedah DropELE menunjukkan ketepatan klasifikasi yang kompetitif dalam set data dunia sebenar, berprestasi baik dalam dua set data dan menunjukkan kepelbagaian yang sebanding atau lebih baik daripada AdaBoost dalam dua set data lain. Model RSO, walaupun bukan yang terunggul, memberikan hasil yang memuaskan berbanding dengan AdaBoost dan secara umumnya mengatasi ensembel RSM, RLO, dan Bagging.

Diversity Injected Ensemble Model for Small Sample-sized Classification Problems

ABSTRACT

Researchers are occasionally required to work with small sample-sized (SSS) data. SSS data tends to undertrain a machine learning algorithm, making it useless due to its extremely small sample size. A low instance-to-feature ratio, or an excessively large number of features relative to a small number of instances, will cause the classification algorithm to overfit in some extreme circumstances of SSS issues. This research proposed two approaches: Random Subspace Oracle (RSO) through the hybridization of Random Subspace Method (RSM) and Random Linear Oracle (RLO) ensembles, and a novel implementation of the dropout approach with an extreme learning machine ensemble (DropELE), both of which focused on solving small sample-sized classification problems. According to the non-parametric Wilcoxon signed-ranks test, the experimental findings showed that the RSO ensemble performed better than a single decision tree and linear discriminant classifier, with a significance level of 0.05. In situations involving other single classifiers, the RSO ensemble's performance was not optimal. The comparison between established ensemble methods was conducted in the experiment that followed. The results showed that the RSO model exhibits comparable performance to the other ensemble approaches, with decision trees as the base classifiers. The DropELE algorithm was proposed as a solution to prevent overfitting difficulties that are frequently encountered in SSS classification problems by reducing the complexity of each base classifier in the ensemble. The research showed that a higher dropout ratio and a properly defined instance-to-feature ratio can greatly improve the performance of the proposed algorithm. According to experimental data, testing accuracy increases significantly from 88.9% to 98% when the dropout ratio is raised from 0.1 to 0.9. It is important to remember that the number of hidden neurons and ensemble size can have a significant impact. In this study, the DropELE algorithm shows that increasing the number of hidden neurons up to a threshold has a positive effect on classification performance. However, beyond this threshold, classification accuracy begins to decline. A similar observation can be made regarding the ensemble size. The performance can be enhanced by increasing the number of base classifiers in the ensemble pool. It is important to acknowledge that the improvement will gradually diminish as more base classifiers are added to the pool. Four real-world medical datasets were used to assess the performance of the two proposed algorithms. The DropELE method shows competitive classification accuracy in real-world datasets, performing well in two datasets and demonstrating comparable or superior diversity to AdaBoost in the other two datasets. The RSO model, while not the top performer, delivers satisfactory results compared to AdaBoost and generally surpasses RSM, RLO, and Bagging ensembles.

CHAPTER 1 : INTRODUCTION

1.1 Background history

Humans gain new knowledge through learning from experience. The ability to learn facts and skills and apply them to adapt to new circumstances is what makes humans intelligent (Legg & Hutter, 2007). Whereas in machine learning, the knowledge gain was done by optimization of mathematical formulations or parameter tuning based on a given database. Machine learning approaches have become mainstream choices in developing practical software such as vision systems, natural language processing, and robot control (Jordan & Mitchell, 2015). The theory behind decision-making in machine learning can be viewed as searching through a large space of viable solutions based on guided training experience, which eventually leads to an optimal solution.

Data is crucial in the field of machine learning, playing an unparalleled significant role in determining the performance and capabilities of models (Rajput et al., 2023). The process of machine learning is based on extracting patterns and insights from data. Data quality, which includes correctness, consistency, and the absence of errors or biases, is of utmost importance, as inferior data might result in inaccuracies or biassed results. Data plays a crucial role in shaping the direction of machine learning projects, requiring careful attention at each step of the modelling process.

The proliferation of advanced technology has made it easier to gather data, resulting in an abundance of extensive datasets and the rise of the "big data" domain (Ren et al., 2016). Nevertheless, certain data cannot be readily obtained due to confidentiality issues, particularly in fields such as healthcare and industry (Kokol et al., 2022). In addition, researchers frequently face conditions characterised by intrinsically restricted data, such as newly identified diseases, rare cancers (Althnian et al., 2021; D. C. Li & Liu, 2012; Ruparel et al., 2013), and investigations necessitating human subjects where recruiting can be costly (Vabalas et al., 2019). In traditional face recognition, it is common to come across small sample sizes because of the differences in instances and features (C. Zhang et al., 2014). Likewise, the initial phases of industrial product development produce limited data. Thus, addressing the challenges of small sample-sized (SSS) data classification remains crucial.

There are various learning schemes in machine learning, the most popular being supervised learning (Hastie et al., 2009). The training data in supervised learning takes the form of ($\boldsymbol{x}$, y), where $\boldsymbol{x}$ is a vector containing discretized features that explain the attributes of an object, and y is the corresponding target output that $\boldsymbol{x}$ should be recognized as. The purpose of supervised machine learning is to predict the output class y^* given by an input $\boldsymbol{x^*}$ through a trained mathematical model. Thus, supervised machine learning is also referred to as classification or pattern recognition in the engineering field (Bishop, 2006).

In machine learning, there are numerous ways to describe the mathematical function used to carry out the learning process. In a common way, it is referred to as the learner because it is the unit that undergoes the learning process. During classification, the module will be referred to as a classifier. An agent will be used when the function is to search through a state space during the optimization process. Also, the term model can also be seen, as those modules are basically a mathematical model of humans' decision-making behavior. This thesis will use all the terms mentioned interchangeably.

Countless number of classification algorithms have been made available in the field of pattern recognition, such as decision tree (TREE), which makes decisions by branching through its tree-like flow structure, the k-nearest neighbour (KNN) looks for nearest reference point(s) as guidance to a final decision, and the Support Vector Machine (SVM) finds an optimal separating hyperplane that maximizes the margin of training data (Burges, 1998; Kowalczyk, 2017). It was known that a single classifier cannot work well across all datasets, thus ensemble methods that combine two or more classifiers were introduced to compensate for the weakness (Dietterich, 2000b).

The basic ensemble model requires two or more classifiers to make up the decision pool (Kumar et al., 2022). The classifiers in the pool are referred to as base classifiers. Base classifiers of the same type will be built under different conditions, such an approach is usually involved in splitting the whole training space into smaller subsets, and each subset will be used to train a classifier (divide-and-conquer). Thus, ensemble methods that follow this principle are said to be homogeneous. Figure 1.1 shows an illustration of a homogeneous ensemble model. Where D_T refers to the overall training data, while $D_{1,2,\dots,n}$, refers to the smaller training subsets partitioned from the overall training data. Each smaller subset will be used to train a classifier. A combiner is used to combine all the predictions from individual base classifiers into one final decision.

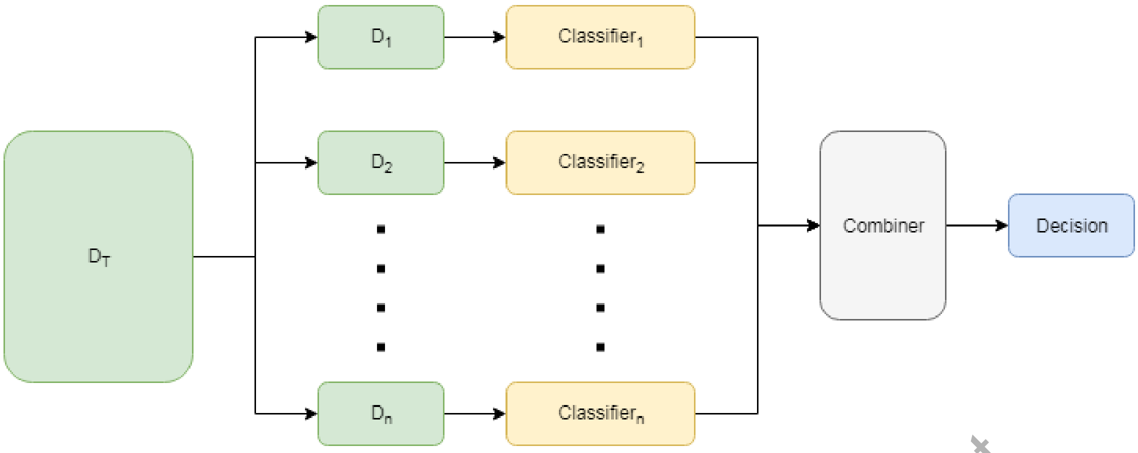

Figure 1.1 : Homogeneous ensemble model.

On the other hand, a heterogeneous ensemble model draws distinct types of classifiers as its base to provide more diverse predictions. This model will have all the base classifiers in an ensemble trained on the whole training space instead of a divide-and-conquer approach. A classifier fusion scheme will be present during the combiner phase to choose the most appropriate base classifier's prediction as the final decision. Figure 1.2 illustrates the heterogeneous ensemble model, with each base classifier coming from a distinct principle. Where the D_T represents the overall training data that all base classifiers will be trained on.

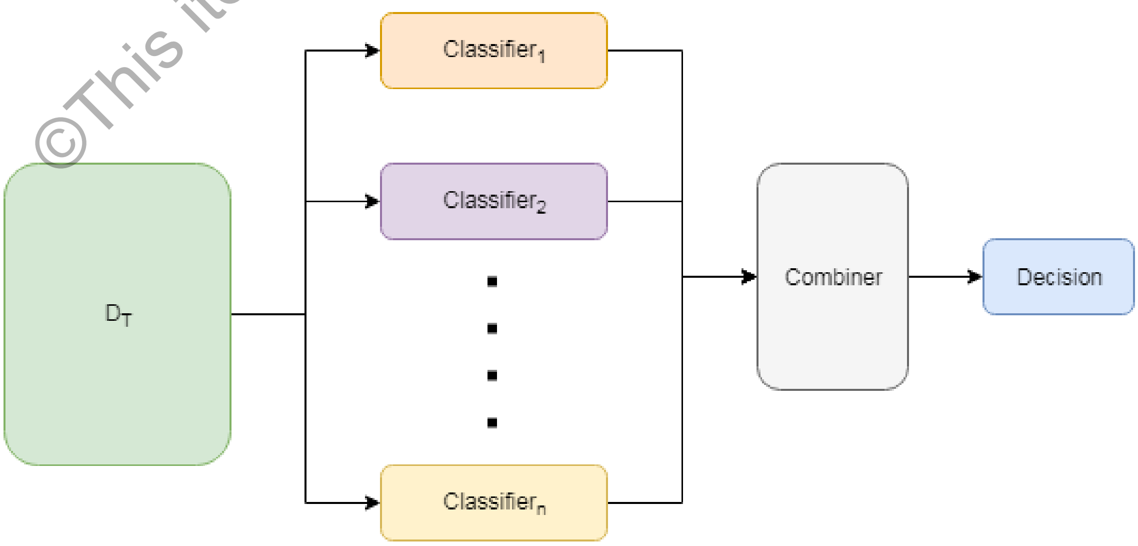

Figure 1.2 : Heterogeneous ensemble model.

Ensemble methods are used for small sample-sized (SSS) data classification (Mian et al., 2024), especially with Bootstrap Aggregating (Bagging) and Random Subspace Methods (RSM). Bagging utilized random sampling with replacement approach, allowing each classifier to train on randomly generated training sets, providing a much more diverse result. A similar approach has been employed by RSM, except that the data selection is done in the feature space. RSM randomly selects a subset of features for the classifier to train on, resulting in several classifiers, each expert in a different feature space. Thus, RSM has been considered for problems with a limited number of samples due to its ability to reduce the data dimension while maintaining the number of training objects (Kuncheva & Plumpton, 2010; Skurichina & Duin, 2002). However, in Zhang, Liang, & Matsuyama (2013), the authors pointed out that the RSM ensemble prioritized global feature extraction rather than local features, resulting in a lack of local discriminant information.

Furthermore, SSS data often leads to algorithm overfitting due to the restricted sample size in relation to the high dimensionality of the data. Overfitting is the phenomenon where the learner excessively fits to the dataset, leading to the creation of an overly complicated model (Montesinos López et al., 2022). An inherent issue with overfitting is that the classifier exhibits superior accuracy on training samples, however its predictive power on previously unseen data is significantly reduced. Apart from manipulating the sample data, dropout regularization is widely used in deep neural network models to prevent the network from overfitting (Dishar & Muhammed, 2023). Although the dropout technique is beneficial for both deep and shallow neural networks, there is limited empirical research that applies the dropout mechanism to shallow networks, particularly in neural networks with random weights (Piotrowski et al., 2020).

1.2 Problem statement

Small sample-sized (SSS) problems are emerging as the focus of current research (T. Zhang et al., 2022). Situations involving SSS data often arise in medical analysis (Alharbi & Vakanski, 2023; An et al., 2023; Tartaglione et al., 2020), machine fault diagnosis (Xie et al., 2022; T. Zhang et al., 2022), and image processing (X. Li et al., 2021; L. Sun et al., 2021). Due to the limited availability of data in these domains, traditional machine learning algorithms struggle to achieve satisfactory performance. The scarcity of data leads to the undertraining of models, making them prone to overfitting and reducing their generalization capabilities. These challenges necessitate the development of novel methods tailored to effectively handle SSS problems.

Recent progress has stimulated the creation of novel algorithms explicitly tailored for SSS classification (Kokol et al., 2022; Rajput et al., 2023; Vabalas et al., 2019; Xu et al., 2023). Nevertheless, there is still a lack of thorough comprehension regarding their performance in comparison to well-established ensemble methods. This limitation arises from the lack of a comprehensive comparison across various datasets. An essential aspect of this analysis involves comparing several ways based on factors such as the number of instances, dimensionality, and instance-to-feature ratio. This enables practitioners to identify the specific strengths and shortcomings of each approach. Through a thorough assessment across several benchmarks, researchers can acquire useful knowledge about the comparative efficiency of current and innovative algorithms, facilitating the creation of more robust models designed for specific situations.