EVOLUTIONARY DEEP BELIEF NETWORK WITH BOOTSTRAP SAMPLING FOR IMBALANCED CLASS DATA

BY

A'INUR A'FIFAH AMRI

A thesis submitted in fulfilment of the requirement for the degree of Master
Computer Science

Kulliyyah of Information and Computer Technology
International Islamic University Malaysia

JANUARY 2019

ABSTRACT

Imbalanced class data is a frequent problem faced in classification task. Imbalanced class occurs when the classes in the dataset has a huge distribution gap between them. The class with the most instances is called the majority class, while the class with the least instances is called the minority class. it caused the result to be skewed towards the majority class instead. Common techniques to overcome or minimize the negative effect of imbalanced class data are data sampling, algorithm modification or hybrid. Deep learning algorithm is a state-of-the-art part of the machine learning algorithms. It is popularized due to better performances when handling complex and high dimensional data. Deep belief network (DBN) is an example of a deep learning algorithm. It is an intricate form of artificial neural network (ANN). It has a deeper layer of neurons that prevents the network from getting stuck when learning from the inputs. It pre-trains the network using Restricted Boltzmann Machine (RBM) and implement backpropagation neural network (BPNN) as a fine-tune step. However, the training time for DBN is longer because of the layers. Also, there is very little apprehension for the exact amount of data or ideal hyperparameters setting to optimize the performance. Due to its complex and deep layers architecture, deep learning needs a lot of training data in order to give good predictions. In this thesis, an optimized DBN is proposed to control the negative outcomes caused by imbalanced class data towards the performance of the algorithm using an evolutionary algorithm. An evolutionary algorithm (EA) is incorporated to provide the optimum dropout number, learning rate, batch size and iteration number of BPNN for fine-tuning in DBN. Bootstrap sampling is also incorporated in the algorithm structure to minimize the bias of data training samples. These modifications improved the ability to predict more accurate outcomes. To evaluate the performance of Evolutionary DBN with bootstrap sampling, an experimental setup involving imbalanced class datasets are conducted. The results of Evolutionary DBN with bootstrap sampling performance is collected and documented in the form of performance metrics. The results are then compared to other machine learning algorithms such as DBN, deep neural network (DNN), BPNN and support vector machine (SVM). According to the outcomes, Evolutionary DBN with bootstrap sampling performed better than DBN and other machine learning algorithms in managing the effects of imbalanced class datasets such as accurate predictions and less partiality. The analysis of statistical tests conducted at the end of this thesis supports the conclusion of the experiment.

الملخص

تعتبر بيانات الفئة غير المتوازنة مشكلة متكررة تواجه عملية تصنيف البيانات. ويحدث عدم التوازن عندما تكون هناك فجوة كبيرة في توزيع البيانات. يُطلق على الفئة التي تحتوي على حالات كثيرة من عدم التوازن فئة الأغلبية ، بينما تسمى الفئة التي تحتوي على حالات اقل فئة الأقلية. وتسبب في انحراف النتيجة لصالح الفئة الاغلبية. هناك تقنيات شائعة للتغلب او الحد من التأثير السلبي لبيانات الفئة غير المتوازنة وهي : اختيار العينات او تعديل الخوارزمية او طريقة الهجين. تعد خوارزمية التعلم العميق قسم من خوارزميات تعلم الاله حيث انها شائعة بسسب الاداء الجيد عند التعامل مع بيانات معقدة وعالية الابعاد. وتعتبر شبكة الوثوق العميق (DBN) أحد امثلة خوارزمية التعلم العميق وهي شكل معقد من شبكة الأعصاب الصناعية(ANN). اذ تحتوي على طبقة اعمق من الخلايا العصبية والتي تحافظ على الاداء الجيد عند التعلم من المدخلات. وتقوم بتدريب مسبق للشبكة باستخدام آلة بولتزمان المقيدة (RBM) والشبكة العصبية العكسية (BPNN) كخطوة ضبط دقيقة. ومع ذلك فأن شبكة الوثوق العميق (DBN) تحتاج الى وقت طويل في عملية التمرين بسبب وجود الخلايا. ايضا تحتاج الى عملية ضبط واعداد مثالي لعناصرها حتى تحصل على النتيجة المطلوبة. كما ان التعلم العميق تحتاج الى بيانات كثيرة من اجل عملية تدريب البيانات لتحصل على توقعات جيدة وذلك بسبب التعقد والعمق في طبقاتها. هذه الأطروحة تقترح تحسين شبكة الوثوق العميق (DBN) للتحكم في النتائج السلبية الناجمة عن بيانات الفئة غير المتوازنة من خلال الخوارزمية التطورية (EA). تقوم هذة الخوارزمية بايجاد القيمة المثلى لكل من : التسرب, معدل التدريب, حجم الدفعة وعدد مرات التكرار. كما انه تم دمج عملية عينات التمهيد في بنية الخوارزمية لتقليل تحيز عينات تدريب البيانات. تهدف هذة التعديلات الى تحسين عملية التنبوء بالنتائج. ومن اجل تقييم التعديلات المقترحة للخوارزمية اجريت تجارب تحتوي على بيانات فئة غير متوازنة. وتم جمع النتائج وتوثيقها على هيئة مقياس اداء. ايضا تم مقارنة نتائج الخوارمزية مع خوارزميات اخرى مثل الشبكة العصبية العميقة (DNN) ونشر شبكة العصبية (BPNN) ودعم شاحنات النقل (SVM) وشبكة الوثوق العميق (DBN). وفقا للنتائج فان شبكة الوثوق العميق (DBN) باندماج مع عينات التمهيد حقق اداء أفضل من شبكة الوثوق العميق (DBN) وبقية خوارزميات التعلم الالي في التحكم بالتأثيرات الناتجة عن بيانات الفئة غير المتوازنة مثل دقة التنبؤ واقل تحيز. وقد تم اجراء تحليل الاختبارات الإحصائية في نهاية الاطروحة حتى يدعم التجارب.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion; it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master Computer Science.

.....
Amelia Ritahani Ismail
Supervisor

I certify that I have read this study and that in my opinion; it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master Computer Science.

.....
Raini Hassan
Internal Examiner

.....
Mohd Zakree Ahmad Nazri
External Examiner

This thesis was submitted to the Department of Computer Science and is accepted as a fulfilment of the requirement for the degree of Master Computer Science.

.....
Raini Hassan
Head, Department of
Computer Science

This thesis was submitted to the Kuliyyah of Information and Computer Technology and is accepted as a fulfilment of the requirement for the degree of Master Computer Science.

.....
Abdul Wahab Abdul Rahman
Dean, Kulliyah of Information
and Computer Technology

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

A'inur A'fifah Amri

Signature:..... Date:.....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

EVOLUTIONARY DEEP BELIEF NETWORK WITH BOOTSTRAP SAMPLING FOR IMBALANCED CLASS DATA

I declare that the copyright holder of this thesis is A'inur A'fifah Amri.

Affirmed by A'inur A'fifah Amri.

....

Signature Date

ACKNOWLEDGEMENTS

I would like to dedicate this piece of work to my parents, sisters, relatives and friends for their emotional support and for making me believe in myself to keep going and complete this thesis. Thank you for reminding me the importance of working hard and working sincerely.

I would like to thank my supervisor, Assoc. Prof. Dr. Amelia Ritahani Ismail for her constant support and guidance. She taught me how to navigate my journey in the academic world which in turn chart my own personal growth. From the exterior, it might seem I've only learned how to read and write journal papers and eventually this thesis, but I also learned to to narrow down my focus for my thesis and discipline myself to finally get the thesis done. Due to her guidance, I've also learned to improve my way of thinking and overcome the challenges in my experiment like a researcher as opposed to merely think like a student. Not only that, I appreciate for her emotional support throughout the process of working on this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 STUDY BACKGROUND

Imbalanced data classification is a classic yet relevant challenge in machine learning. It deters the optimal prediction of the dataset, which can affect negatively to decision making. When data instances are too expensive to be collected or the data are simply scarce, it emanates data disparity between the categorical classes.

The class with the most instances is called the majority class, while the class with the least instances is called the minority class. Imbalanced data problems usually occurs because the data is unavailable and the attempt to retain it is expensive (Berry et al., 2012). Since machine learning rely on training data to analyse and predict the outcome of the models, it find difficulties when dealing with imbalanced class datasets.

Common outcomes from this issue are misclassification and fluctuating error rates. This pose an issue when the minority class is the sought after prediction as these scenarios does not happen on regular basis. Since imbalanced class problem is a frequent obstacle, there has been many approaches formulated to encounter such complication. Imbalanced class problem can either be solved using data sampling method, algorithm modelling method or both.

Data sampling method involves in tweaking the data itself such as undersampling and oversampling. Undersampling is the process of removing selected instances from the majority class in the dataset. Undersampling might dispose certain instances that can be important for an algorithm model to learn from (Y. Liu et al., 2010). Oversampling is the process of duplicating the instances from the minority class so that it

has the same amount as the instances in the dataset. However, this method can produce a negative effect commonly overfitting (Y. Liu et al., 2010).

Another approach to overcome the repercussions of imbalanced class data is by algorithm modelling. This includes modelling an algorithm hybrid that solves the issues of specific data input. Deep learning algorithms has gained many interests as a state-of-the-art machine learning approaches. Deep learning algorithms has proven to able produce high accuracy results as well as extracting high level abstraction of many domains(Wang et al., 2012; Mohamed, Dahl, & Hinton, 2012; Le & Provost, 2013). Some examples of deep learning algorithms are deep belief network (DBN), deep neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN) and convolutional deep belief network (CDBN). DBN is a stacked of neurons made up of restricted Boltzmann machine (RBM). The network is fine tuned using backpropagation neural network (BPNN) approach. DBN is widely researched algorithm and is proven to produce good results in high abstract domains such as emotion recognition (Le & Provost, 2013) and acoustic modelling (Mohamed, Hinton, & Penn, 2012).

Evolutionary algorithm (EA) acts as an optimization algorithm. It is inspired by biological evolution process. It involves the process of selection, crossover and recombination that allows improvement in selected aspects of an algorithm. There many types of EA, such as genetic algorithm (GA), particle swarm optimization (PSO)(Y. Zhang et al., 2014) and whale optimization algorithm (WOA)(Mirjalilia & Lewis, 2016). GA is a commonly used EA because it yields good results when utilized as optimization algorithm.

1.2 PROBLEM STATEMENT

Imbalanced class can affect negatively in decision making process by providing poor results due to misclassification and fluctuating error rates (Weiss & Provost, 2001). Common methods used to overcome such problem is by data sampling or algorithm modelling (Zhai et al., 2015). Data sampling method is commonly used and is useful when it comes to handling imbalanced class issue because it deals the problem directly. The basic approach that is frequently used is undersampling, oversampling or a hybrid of both. However, the issue with undersampling is that it is possible to get rid of crucial data needed for prediction (H. Han et al., 2005), while the issue with oversampling is that it causes overfitting in learning (Y. Liu et al., 2010). Nevertheless, implementing a data sampling method is still sought after since it can minimizes the negative effects of imbalanced class problems because it deals with the data directly.

Another plausible solution for imbalanced class problem is by algorithm modelling. Deep learning has shown promising results in many domains, especially the ones that require high level abstraction and has complex data features, such as image processing, emotion detection and handwriting recognition (Wang et al., 2012; Mohamed, Dahl, & Hinton, 2012; Le & Provost, 2013). An example of deep learning algorithms is deep belief network (DBN). DBN can learn from complex feature input such as emotion recognition (Le & Provost, 2013) and acoustic modelling (Mohamed, Hinton, & Penn, 2012). Therefore, it can learn the features from an imbalanced class dataset and classify it correctly. Despite the promising performance of DBN in various fields, the algorithm is generally computationally expensive and unable to achieve competent result when learning from inadequate amount of data (Le & Provost, 2013; Mohamed, Hinton, & Penn, 2012).

1.3 RESEARCH HYPOTHESES

Below is the hypothesis of the study:

H_1: The Evolutionary DBN with bootstrap sampling has shown high performance in terms of performance metrics and statistical analysis as compared to other deep learning and machine learning algorithms when handling imbalanced class datasets.

1.4 RESEARCH OBJECTIVES

This study embarks on the following objectives:

1. To study the effects of DBN algorithm with imbalanced class datasets.
2. To propose an optimized DBN model using an evolutionary algorithm with bootstrap sampling for imbalanced class datasets.
3. To compare the classification performance of optimized DBN algorithm with other machine learning algorithms for imbalanced class datasets.
4. To analyze the performance of optimized DBN using statistical techniques.

1.5 RESEARCH QUESTIONS

This research aims to answer the following questions:

1. How does the imbalanced class datasets affect the performance of DBN algorithm?
2. How to optimize the DBN algorithm?
3. Does the implementation of an evolutionary algorithm and bootstrap sampling on DBN increases the performance when imbalanced class datasets are used?

1.6 SIGNIFICANCE AND CONTRIBUTION

The major contribution of this research is to propose an Evolutionary DBN with bootstrap sampling for imbalanced class problems. This optimized DBN algorithm is helpful for predicting imbalanced class datasets of binomial category.

1.7 THESIS STRUCTURE

This thesis is organized into 6 chapters. Chapter 2 presents the literature review studied for this thesis. The literature review section studies the background of imbalanced class challenges in datasets and deep learning algorithms. Machine learning and evolutionary algorithms are also reviewed for comparison. Chapter 3 provides the research methodology of this thesis. It describes the steps taken to introduce an optimized DBN for imbalanced class data classification.

In Chapter 4, the experimental setup of the research is explained in details. The imbalanced class datasets, algorithms, performance metrics and statistical tests utilized for the experiment is elucidated. Chapter 5 presents the results and performance of Evolutionary DBN with bootstrap sampling and compares it with other deep learning and machine learning algorithms. Finally, Chapter 6 is dedicated to the conclusion of the experiment and future works recommendation based on the result analysis.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This section presents the literature review done for imbalanced class problem, machine learning, deep learning and evolutionary algorithms. Section 2.2 describes imbalanced class affair in datasets and the usual techniques employed to overcome or minimize its negative effects towards algorithm prediction result. Sections 2.4 and 2.5 explains the details of machine learning and deep learning algorithms in general and their involvement with imbalanced class datasets. Section 2.6 elucidates evolutionary algorithm and its examples when dealing imbalanced class datasets.

2.2 IMBALANCED CLASS DATASET

Imbalanced class ordeal in a dataset is a common classification task problem. According to Hensman & Masko (2015) and Yan et al. (2015), imbalanced class refers to the *"disparity of data dispensation between the classes"*. The distribution of instances between the classes are not balanced. This affects the performance of an algorithm when doing prediction of the classes (C. Zhang et al., 2018). The class that has more training values is called the majority class and the class that has the least or most missing data values are called the minority class (Swersky et al., 2010).

Minority data class is a realistic problem that the real-world situation faced (C. Zhang et al., 2018) because most of the time even for an important dataset such as cancer detection (N. V. Chawla et al., 2004) and bank fraud(Awoyemi et al., 2018), the data instances are scarce. It can be expensive if the new data needs labelling (Berry

et al., 2012). Unfortunately, most of the algorithms that showed stable and promising performance when using balanced data in classification tasks displayed conflicting outcome when imbalanced class dataset is used (Ghahabi & Hernando, 2014). Prediction of minority class is presumed to have a higher error rate compared to the majority class and its test examples are often wrongly classified as well (Weiss & Provost, 2001). Imbalanced data distribution among the classes causes deficient classification models (C. Zhang et al., 2018). The algorithm that performs on balanced dataset will not perform as good when using an imbalanced dataset (C. Zhang & Tan, 2016), regardless how good the model is.

In a study done by Yan et al. (2015), an imbalanced class dataset in multimedia format is implemented as the input for CNN. The dataset is a TRECVID dataset, which means it is in the form of video. The outcome shows that the error rate fluctuate unlike when using a balanced dataset, the error rate of the algorithm decrease steadily.

There are a few commonly used methods utilized to tackle the challenges of imbalanced class dataset. The first method is using machine learning algorithm and model hybrids according to the input types (W. Liu & Chaw, 2011; Zhai et al., 2015). Another method is by data preprocessing of the imbalanced dataset itself (Zhai et al., 2015). Both data sampling and machine learning approaches are explained in Section 2.3 and Section 2.4 respectively.

2.3 DATA SAMPLING METHOD

Data sampling is a commonly used method to rebalance the data instances (W. Liu & Chaw, 2011). This way, the negative effects of imbalanced class data is handled by directly manipulating the amount of instances in the dataset. In this section, the approaches that are discussed are oversampling, undersampling and bootstrap sampling.

Fernandez et al. (2011) highlights that this method can liberate the performance of an algorithm to be affected.

2.3.1 Oversampling

Oversampling is when the minority class in the dataset are duplicated until it has the same amount or as many as the majority classes in the dataset (Y. Liu et al., 2010; Ganganwar, 2012; W. Liu & Chaw, 2011). Even though it seem ideal for training algorithm because the dataset will have more instances, the disadvantage of an oversampling is that it can cause the classifier the problem of overfitting because the algorithm will learn the redundant values all over again (Y. Liu et al., 2010). "Synthetic Minority Oversampling Technique" (SMOTE) is a common algorithm to optimize the use of oversampling technique (Y. Liu et al., 2010; Fernandez et al., 2011; N. Chawla et al., 2002) . However, there are many variation of SMOTE to suit each dataset (W. Liu & Chaw, 2011). Hensman & Masko (2015) performed an oversampling technique to minimize the effects on convolutional neural network (CNN)

2.3.2 Undersampling

Undersampling is when the majority class in the dataset is downsized, which means the values are removed in order to have the similar amount as the minority data class (Y. Liu et al., 2010; Ganganwar, 2012; Fernandez et al., 2011). If the minority class is desired, the majority class of the dataset is probably unaffected considering the algorithm can focus on the balanced amount of instances instead. However, the downside of an undersample is that there is a possibility that the important values are taken away from the majority class which can result in an inaccuracy (Y. Liu et al., 2010; H. Han et al., 2005; C. Zhang & Tan, 2016). According to Y. Liu et al. (2010), undersampling is considered

superior to oversampling in terms of avoiding overfitting.

2.3.3 Bootstrap Sampling

Bootstrap sampling is when a small sample is derived from its original sample iteratively (Yan et al., 2015; Rosca, 2014). This method basically reuse its training samples and this is a suitable technique to avoid data redundancy as well as data disposal. Megumi et al. (2015) conducted a neuroscience experiment involving fMRI neurofeedback. Bootstrap sampling was utilized as a method to assess the experiment's difference in correlation between the neurofeedback and other networks.

Bootstrapping sampling is a frequently adopted technique implemented to improve performance of deep learning algorithms with imbalanced class data (Y. Liu et al., 2010; Berry et al., 2012). Yan et al. (2015) implemented convolutional neural network (CNN) to classify an imbalanced multimedia dataset. Bootstrap sampling method is integrated with the algorithm to minimize its fluctuating error rate. The experiment yielded high F1-score as compared to another framework proposed by Tokyo Institute of Technology (TiTech).

In another literature, Berry et al. (2012) implemented bootstrap sampling as a method to improve both computational time and accuracy rate after training the imbalanced and unlabeled data using deep belief network (DBN). The result is recorded to have 41% decrease of error rate that needs human intervention as compared to no bootstrapping implementation. Z. Sun et al. (2018) predicts wind speed and wind power using deep belief network and optimized random forest. The experiment has inconsistent amount of data because some data are simply unavailable. Therefore, the experiment employed bootstrap sampling as an approach to resample the training data to improve the performance of their model.

2.4 MACHINE LEARNING ALGORITHM

Machine learning is a field in Artificial Intelligence (AI). It focuses on algorithms that are prototyped from natural intelligence such as the human brain or animal interactions so that it has the ability to learn. Machine learning algorithms are categorised into supervised learning, unsupervised learning and reinforcement learning. Supervised learning is when the input data is labeled and the output needed is classification, recognition or prediction. Unsupervised learning is when the input data available are not labeled and the output often involves clustering.

A few examples of supervised learning algorithms are artificial neural networks (ANN), support vector machine (SVM), naive Bayes, decision trees and k-Nearest Neighbour (k-NN). Although machine learning algorithms are essentially stable and have strong mathematical and statistical basis, it is not robust when applied to different domains since it lacks the domain knowledge and data processing. There are few machine learning algorithms examples that are used to tackle imbalanced dataset classification. In this section, a few examples of the algorithms and the result towards the imbalanced dataset classification will be explained.

The author Y. Liu et al. (2010) used support vector machine (SVM) as the main algorithm and showed that the effect of data disparity results in a *"high false negative rate"*. Zou et al. (2008) applied a Genetic algorithm (GA) sampling to an imbalanced protein data for prediction and continue classifying it with SVM. Another paper, W. Liu & Chaw (2011) modified k-nearest neighbours (kNN) algorithms to counter the effect of imbalanced class dataset to the algorithm.

This section will present the related work conducted in respect to imbalanced dataset classification using ANN and SVM. Relevant works of applying each algorithm

as a method of countering the problems of imbalanced data is presented in each sections.

2.4.1 Backpropagation Neural Network Algorithm

Artificial neural networks (ANN) are modelled after the human brain networks (D. Zhang & Xu, 2014; Amato et al., 2013). It is widely known used for supervised learning and recognising patterns from input data set by weight adjustments (D. Zhang & Xu, 2014). Common examples of ANN are feed forward, radial basis function (RBF) and Kohonen's Self Organizing Map, to name a few. ANN's ability to scrutinize nonlinear data and to design complex models has allow it to be applied in studies of different fields (Amato et al., 2013; Drew & Monson, 2000).

The most common neural network algorithm used is the backpropagation neural network (BPNN). BPNN has three layers, which consists of an input layer, a hidden layer and an output layer (D. Zhang & Xu, 2014). The layers are made up of interconnected nodes by a weighted association and the number of nodes of the layers depend on the problem domain (Amato et al., 2013). The input layer will accept the data for training or testing and pass the weights to the connected hidden layer. Hidden layer can be one or more and it will continue calculating the weights it received and send it to the output layer where the result is produced. BPNN compares its real output and target output and adjust its weight according to the error and propagates back to its network. However, BPNN is commonly known to have the problem of overfitting when learning (Lanbouri & Achchab, 2015).

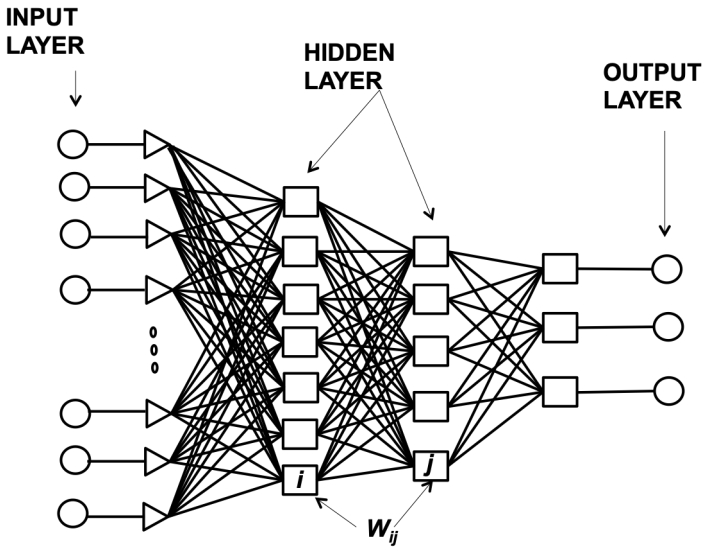

Figure 2.1: An example of a neural network with two hidden layers (Amato et al., 2013).

Arora et al. (2010) implemented back propagation neural network classifier to categorize Devnagari handwritten classes and compared its performance with SVM using the same handwritten data set. The experiment result for BPNN performance is 90.44% for testing data set accuracy. Another work by J.Pradeep et al. (2011) proposed a diagonal based feature extraction and used a "feed forward backpropagation" neural network to classify the data based on the new feature extraction. The experiment achieved 96.52% with 54 features and 97.84% with 69 features accuracy rate.

In tackling imbalanced data, Cao et al. (2013) presented a cost sensitive back-propagation neural network for a multiclass imbalanced data, as opposed to the "limited" binary class imbalanced data. D. Zhang & Xu (2014) implemented BPNN as a method to credit scoring. The dataset is from "The German Credit Dataset" and is imbalanced. The dataset is meant to separate between eligible credit applicants. Other than that, BPNN is also employed to predict protein disorder and manage to get accuracy rate of 91.00% on average of 4 BPNN models (Oh, 2013). The dataset is imbalanced and is procured from "Korea Institute for Advanced Study"

2.4.2 Support Vector Machine Algorithm

According to Arora et al. (2010), SVM can be defined as a "binary classifier", where the outcome will be divided into two groups based on the optimum hyperplane.

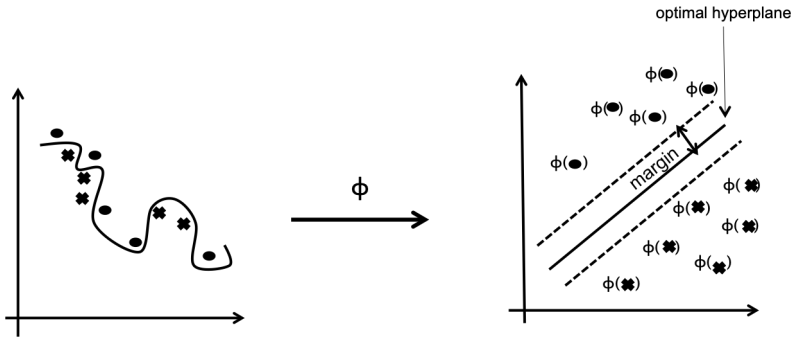

Figure 2.2: SVM mapping nonlinear problem to linear using optimum hyperplane (Ren, 2012).

Niu & Suen (2011) implemented a hybrid of SVM and CNN for classifying MNIST handwritten digits dataset. Feature extraction is done using CNN and SVM acts as a "recognizer". (Arora et al., 2010) compared the performance of SVM and ANN using the Devnagari handwritten recognition problem. SVM performance in the experiment achieved 92.38% for testing accuracy.

In countering the imbalanced data classification problem using SVM, its weight and activation function are manipulated in order to increase the classification accuracy (Hwang et al., 2011). Tang et al. (2009) stated that SVM outperforms other conventional classifiers when a moderate imbalanced data is used. Even so, when a high imbalanced data is used instead, SVM classifier can still produce a biased result. Most works using SVM to counter imbalanced data only focused on the performance and not efficiency, hence, SVM can be a slow classifier (Tang et al., 2009). However, Zou et al. (2008) stated that SVM could not perform imbalanced data classification successfully based on