Faculty of Information and Communication Technology

HYBRID NEURAL NETWORK WITH K-MEANS FOR FORECASTING RESPONSE CANDIDATE IN DIRECT MARKETING

Ramadhan Rakhmat Sani

Master of Computer Science (Software Engineering and Intelligence)

2014

BORANG PENGESAHAN STATUS THESIS*

JUDUL : HYBRID NEURAL NETWORK WITH K-MEANS FOR FORECASTING RESPONSE CANDIDATE IN DIRECT MARKETING
SESI PENGAJIAN : 2013 - 2014
Saya : RAMADHAN RAKHMAT SANI

(HURUF BESAR)

Mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hak milik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

_____ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

_____ TIDAK TERHAD

_____ _____

(TANDA TANGAN PENULIS) (TANDA TANGAN PENYELIA)

Alamat Tetap: _____ *Prof. Madya Dr. Burhanuddin Mohd. Aboobaider*

_____ Nama Penyelia

_____

Tarikh: _____ Tarikh: _____

CATATAN : * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana (PS).
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

HYBRID NEURAL NETWORK WITH K-MEANS FOR FORECASTING RESPONSE CANDIDATE IN DIRECT MARKETING

RAMADHAN RAKHMAT SANI

A thesis submitted

In fulfillment of the requirements for the degree of

Master of Computer Science (Software Engineering and Intelligence)

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2014

DECLARATION

I declare that this master project entitled "*Hybrid Neural Network With K-Means for Forecasting Response Candidate in Direct Marketing*" is the result of my own research except as cited in the references. This master project has not been accepted for any degree and is not currently submitted in candidature of any other degree.

Signature : ……..

Name : Ramadhan Rakhmat Sani

Date : January 2014

APPROVAL

I hereby declare that I have read through this project report and in my opinion this project report is sufficient in terms of scope and quality for the award of the degree of Master of Computer Science (Software Engineering and Intelligence).

Signature : ……...

Name : Prof. Madya Dr. Burhanuddin Mohd. Aboobaider

Date : January 2014

DEDICATION

To Allah SWT that always gives a grace and guidance in my life

To my family, who always love and pray for me, give me support to keep fight in study

To my beloved who always remind me and support all these time

To my friends who fought together and always encourage each other

ACKNOWLEDGEMENT

Alhamdulillah, all praises to Allah, for the strengths and the blessing in completing this project entitled: *Hybrid Neural Network With K-Means For Forecasting Response Candidate In Direct Marketing with imbalanced data.*

Special appreciation goes to my supervisor, Prof. Madya Dr. Burhanuddin Mohd. Aboobaider, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the experimental and project works have contributed to the success of this research.

I would like to express my appreciation to Director of International Office Prof. Dr. Nanna Suryana Herman, for their support and help towards my postgraduate affairs.

I am especially grateful to University of Dian Nuswantoro (UDINUS) for the opportunity given to me to continue my study and their kindness in giving financial support during my study here. Special thanks also go to Dr. Ir. Edi Noersasongko, Dr. Abdul Syukur, Dr. Kusni Ingsih, and others for their support.

Sincere thanks to all my friends and senior, especially Mr. Affandy, Mr. Sriyanto, Herdi, Hayati, Fatma, Andita, Egia, Danang, Prajanto, mobility student from UDINUS, PPI UTeM members, and others for their kindness and moral support. Thanks for sharing their experience and knowledge, for their friendship and deep memories during study in Melacca.

Finally, the deepest gratitude goes to my family, especially my blessed parents, Mr. Subekti and Mrs. Wiwiek Churiyatiningtyas, my sister and brother, and my beloved Amilia Fitri Utami for being to give me uncountable love, prayer, and encouragement.

ABSTRACT

The larger Bank‘s electronic data customer provides difficulty a marketing campaign. An efficient marketing campaign is needed to promote a product and services. The predictive data mining techniques use to help a marketing analyst provide more value to their customers by the right offer because of decreasing in responses to a direct marketing campaign. Distribution of customer data record in marketing response data are often found issue of imbalanced dataset. This study proposed hybrid Neural Network (NN) methods in data mining to support direct marketing analysis and forecast. Backpropagation NN is supervised learning methods that analyze data and recognize to solve many problems in the real world by building a model that is trained to perform well in some non-linear problems. K-means algorithm grouping process by minimizing the distance between the data and designed can handle very large dataset also continuous and categorical variable for handling imbalanced dataset. This research concerns on binary classification which is classified into two classes. Those classes are yes and no. The data was collected from the Machine Learning Repository Dataset in the University of California Irvine (UCI).This experiment compares hybrid K-Means + NN with basic NN. The result shows the improvement of accuracy from 91.53% to 91.59%, recall 22.15% to 27.7% and F-Measure 44.23% but not to precision from 61.69% to 60.75%.

ABSTRAK

Data elektronik pelanggan yang lebih besar daripada bank yang menyediakan kesukaran kempen pemasaran. Satu kempen pemasaran yang cekap diperlukan untuk mempromosikan produk dan perkhidmatan. Teknik-teknik perlombongan data ramalan gunakan untuk membantu seorang penganalisis pemasaran memberikan nilai kepada pelanggan mereka dengan tawaran yang tepat kerana berkurangan di balas kepada kempen pemasaran langsung. Pengagihan rekod data pelanggan dalam data sambutan pemasaran sering dijumpai isu dataset seimbang. Kajian ini dicadangkan Rangkaian Neural hibrid (NN) kaedah dalam perlombongan data untuk menyokong analisis pemasaran langsung dan ramalan . Rambatan balik NN diselia kaedah pembelajaran yang menganalisis data dan mengiktiraf untuk menyelesaikan banyak masalah dalam dunia sebenar dengan membina model yang dilatih untuk prestasi yang baik dalam beberapa masalah yang tidak linear. K-Means algoritma proses perkumpulan dengan meminimumkan jarak di antara data dan direka boleh mengendalikan dataset yang sangat besar juga berubah-ubah selanjar dan untuk mengendalikan dataset seimbang. Ini kebimbangan penyelidikan klasifikasi binari yang diklasifikasikan kepada dua kelas. Kelas-kelas adalah ya dan tidak. Data dikumpulkan dari Mesin Pembelajaran Repository dataset di Universiti California Irvine (UCI). Eksperimen ini membandingkan hibrid K-Means + NN dengan asas NN. Hasilnya menunjukkan peningkatan accuracy dari 91,53 % kepada 91,59 % , recall 22.15 % kepada 27.7 % dan F- Measure 44,23 % tetapi tidak untuk precision dari 61,69 % kepada 60,75 %.

TABLE OF CONTENT

DECLARATION i
APPROVAL ii
DEDICATION iii
ACKNOWLEDGEMENT iv
ABSTRACT v
ABSTRAK vi
TABLE OF CONTENT vii
LIST OF ABBREVIATIONS x
LIST OF FIGURES xi
LIST OF TABLES xii
CHAPTER
1. INTRODUCTION 1
1.1 Introduction 1
1.2 Background of Study 3
1.3 Problem Statements 4
1.4 Research Question 5
1.5 Research Objective 5
1.6 Scope of Study 5
1.7 Significant of Study 6
1.8 Project Report Overview 6
1.9 Chapter Summary 7

2. LITERATURE REVIEW 8
2.1 Introduction 8
2.2 Direct Marketing 8
2.3 Research Trend on Direct Marketing 9
2.4 Data Mining 12
2.4.1 Definition of Data Mining 12
2.4.2 Data Mining Techniques and Applications 13
2.4.3 Classification and Prediction in Data Mining 13
2.4.4 Clustering in Data Mining 14
2.4.5 ROC Curve (Receiver Operating Characteristic) 16
2.5 Neural Network 18
2.5.1 Advantages and Disadvantages of Neural Network 18

2.5.2 Backpropagation 19
2.6 Feature Selection 25
2.7 Class Imbalance Problem 26
2.8 Chapter Summary 27

3. METHODOLOGY 30
3.1 Introduction 30
3.2 Type of Research Method 30
3.3 Research Design 31
3.4 Data Collection 32
3.5 Preprocessing data 34
3.6 Proposed Method 34
3.7 Performance Measurement 38
3.8 Research Tool 40
3.9 Chapter Summary 41

4. EXPERIMENTAL AND RESULT 42
4.1 Introduction 42
4.2 Data Preparation 42
4.3 Data Reduction And Discretization 44
4.4 Experiment Result 45
4.3.1 The Comparison Result of Neural Network With Determination of Hidden Layer 45
4.3.2 The Comparison Result of Neural Network With Determination of Training Cycles 47
4.3.3 The Comparison Result of Neural Network With Determination of Learning Rate 49
4.3.4 The Comparison Result of Neural Network With Determination of Momentum 50
4.3.5 The Comparison Result For Basic Neural Network And K-Means-NN With Determination Cluster 52
4.5 Discussion Result 54
4.6 Chapter Summary 55

5. CONCLUSION 56
5.1 Introduction 56
5.2 Research Background and Propose Of Study 56
5.3 Literature Review 56

5.4 Proposed Method 57
5.5 Experiment Result 57
5.6 Conclusion 58
5.7 Future Work 58

REFERENCES **59**

Appendix A Literature Reviews Of Direct Marketing 63
Appendix B Data Set 67
Appendix C Modeling In Rapidminer 75
C1 Main Process of K-Means + NN Model 76
C2 Neural Network Model 77
C3 Improved Neural Network 78
C4 Performance of BASIC NN Model 79
C5 Performance of K-Means + NN Model 80

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
BNS	-	Bayesian Network
BP	-	Backpropagation
CART	-	Classification and Regression Tree
EP	-	Evolutionary Programming
GA	-	Genetic Algorithm
kNN	-	K-Nearest Neighbors
KDD	-	Knowledge Discovery and Data Mining
MLP	-	Multilayer Perceptron
MVC	-	Model View Controller
UCI	-	University of California Irvine
SOM	-	Self Organizing Map
SVM	-	Support Vector Machine
TN	-	True Negative
TP	-	True Positive
Imbalanced dataset	-	Means if one class contains significantly proper handling more sample than the other

LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1	Classification and Prediction in Data Mining	14
2.2	K-Means Clustering Algorithm	15
2.3	Roc Graphic (Discrete/Continuous Case)	17
2.4	BP Neural Network with Three Layers of Units	20
2.5	Multilayer Feed-Forward Neural Network	22
3.1	Research Design	31
3.2	The Proposed Method	35
3.3	Screenshot of Rapidminer	41
4.1	Atibure Weighting	43
4.2	Accuracy Based on Hidden Layer	47
4.3	Accuracy Based On Training Cycles	48
4.4	Accuracy Base on Learning Rate	50
4.5	Accuracy Based on Momentum	51
4.6	Comparison Result Of Basic NN And K-Means + NN	53
4.7	Comparison Area Under Roc (AUC) Result Basic NN And K-Means+NN	53

LIST OF TABLES

TABLE	TITLE	PAGE
2.1	Initial Input, Weight and Bias Values	22
2.2	Value of Weight	24
2.3	Comparison Using Neural Network	28
3.1	Dataset of Term Deposit Between July 2008 And June 2009	33
3.2	Table of Confusion Matrix (Vercellis, 2009)	38
4.1	Weight by Information Gain Ratio	43
4.2	The Attributes are Used	44
4.3	Experiment Determination Of Hidden Layer	46
4.4	Experiment Determination Of Training Cycles	48
4.5	Experiment Determination of Learning Rate	49
4.6	Experiment Determination of Momentum	51
4.7	Comparison Determination of Value K	52
4.8	Measurements Before Using K-Means and After Using K-Means	54

CHAPTER 1

INTRODUCTION

1.1 Introduction

Due to large customers in company, an efficient marketing campaign is needed to promote a product and services. The large competitions to promote a new product encourage selecting an efficiency campaign. In the current financial crisis is a major challenge for banks to increase financial assets. The strategies adopted by offering attractive interest rates for term deposit application is good, especially with the direction of using the marketing campaign for the reduction of costs and time with customers that respond statistical products through direct marketing. The customer who respond the product are less than 8 % of 79354 customers (Moro et al., 2011). Thus, there is a need to improve the efficiency of the contact but do not reduce the number of successes obtained.

Currently mass marketing is no longer a method that an efficient and reliable. Displacement of traditional marketing to direct marketing make a reasonable grounds to selecting customer base on necessity and characteristic for their target promotion. Direct marketing is different from mass marketing that targets general public by using media (Bose and Chen, 2009; Patil et al., 2009).

The latest development in information technology and the increasing confidence placed in complex computer systems, several direct marketing strategies have been used by Web sites and newsletters, including campaign strategy implemented in collaboration with

a group of public relations to increase media coverage, development and utilization Speakers Bureau (Szymanski, 2012).

The customer was asked to write down their responses over time are then stored in the database to estimate the number of customers who will eventually respond in direct marketing (Chun, 2012). The predictive data mining techniques use to help an marketing analyst provide more value to their customers by the right offer because of decreasing in responses to a direct marketing campaign (Breur, 2007). Classification is made more interesting by the fact that today's marketing environment as well as save on the amount outstanding customer information with a very low cost, including socio-demographic, transactional purchasing behavior, attitude data (Coussement and Buckinx, 2011). We can estimate the expected number of responses or the overall response rate, and use that information in making important decisions which registered in customer‘s response. The trends of sequential and time series still be open important problem in data mining for clustering, classification and prediction the trends of these data (Yang and Wu, 2006).

The larger customer‘s electronic data from a bank provides difficulty a marketing analyst to make a decision strategy for marketing campaign (Elsalamony and Elsayad, 2013). For overcome rising costs and declining response rates from customers, direct marketing using predictive models to analyze customer data (demographic and historical purchasing data) to select customers who are more likely to respond to promotions that provide a higher response rate and it is an effective method for marketing (Sing‘oei and Wang, 2013). Among the the different domain of marketing, customers segmentation or profiles recognized as the essential area. Being customer centric based on marketing paradigm in targeted marketing make unsolicited marketing is costly and ineffective. Along with these reasons, there is an increasing effort to collect and analyze customer data for better marketing decisions (Olson and Chae, 2012).

From direct marketing campaigns that have been made, the necessary targeting offers to customers for increase the economic benefits of an enterprise, either by acquiring new customers or to generate additional revenue from existing customers (Talla Nobibon et al., 2011).

1.2 Background of Study

The fundamentally of important role in the company is customers; organizations are always competing to explore the potential of the best to offer to them. Data mining model provides support in performance of these campaigns. Different Methods classifications can be combined to improve the accuracy of the model. After the successful implementation of the function classification / prediction, or as a stand-alone Data Mining function, optimization can be performed to determine the factor settings (or design parameter) that produce the desired response (Köksal et al., 2011).

In another study aimed to verify the effectiveness of direct mail marketing campaigns by using ANNs and, in particular, MLP ANN. Identify characteristics and purchase constitute intentions of targeted campaigns. In direct marketing campaign to try to maximize profits as a positive response from customers and prospects list (Guido et al., 2011). Neural network has advantages in non-linear prediction, has a very good performance in parallel processing and the ability to tolerate faults.

It is very appropriate to the characteristics of the dataset forecasting term deposit in this study. Neural network is a method often used to predict the costumer on direct marketing because the data is presented for this method to be large and non-linear (Gill, 2005). The most popular technique in neural network method is backpropation algorithms are widely used to solve many problems in the real world by building a model that is

trained to perform well in some non-linear problems (Park, Lee, & Choi, 2009). We can also use the clustering technique with K-Means algorithm to handle the imbalanced dataset. This combination is considered to be a part of the research contribution.

1.3 Problem Statements

The larger customer‘s data from a bank for marketing campaign requires time consuming, expensive cost, inconvenient analysis and it‘s not efficient for target marketing (Elsalamony and Elsayad, 2013).

The percentage response rate of customers who actually buy the product are typically very low, less than 1 % (Moro et al., 2011; Patil et al., 2009). Identifying costumers who are likely to respond to new offers will be a difficult task through manual perusal of a large customer database (Festus Ayetiran and Barnabas Adeyemo, 2012). The complexity of the data mining models that make it difficult for marketers to use and understand to knowledge on data mining skills (Sing‘oei and Wang, 2013). In most cases, ANN classifiers perform poorly to handle imbalanced data because their design for balance dataset (Adam et al., 2012). However, class imbalance problem using sampling techniques have their drawback (Daneshmandi and Ahmadzadeh, 2013).

Although many studies have provided important insights into direct marketing, but interest in understanding the importance of this issue still lacking. Organizations are now realizing the importance of data mining in their strategic planning to get hidden predictive information from large amounts of data.

1.4 Research Question

Based on the background of the study and the problem statement, the research questions in this study are follows:

1.4.1 How to improve an efficient marketing campaign?

1.4.2 What is the suitable method to handle class imbalance dataset?

1.4.3 How to get better accuracy of ANN for marketing responses dataset?

1.5 Research Objective

This study has three objective based on research questions. Those objectives are:

1.5.1 To increase the effectiveness marketing campaign using data mining technique.

1.5.2 To investigates an appropriate prediction model in order to address class imbalanced problem.

1.5.3 To evaluate an experimental marketing using ANN.

1.5.4 To analyst the impact of reducing class imbalanced using K-mean algorithm.

1.6 Scope of Study

The scope of this study includes the following:

1.6.1 The research study uses a dataset from University of California Irvine (UCI) machine learning repository dataset.

1.6.2 The measurement of class imbalanced data used accuracy, precision, recall, F-Measure and ROC Curve.

1.6.3 Experiments were performed using RapidMiner Software.

1.7 Significant of Study

Many researches related to do data mining response in direct marketing have been conducted recently.

1.7.1 The use of ANN is to improve the predictive classification accuracy rate in term of marketing analysis.

1.7.2 The propose K-Means algorithm provides to reduce class imbalanced dataset

1.8 Project Report Overview

This study provide three chapter of this project report. The report structure is as follow:

Chapter 1

Chapter 1 is the introduction part of this study. This chapter provides information about the origins of the research. A brief outlining of The Background of Study, Problem Statements, followed by Research Questions, Research Objectives, Scopes of Study, Significant of Study and as well as the Project Report Overview and Summary.

Chapter 2

Chapter 2 is the literature review part. This section describes about Direct Marketing, Trend Research of This Study, Data Mining, Neural Network, Class Imbalance Problem, Feature Selection and Summary.

.

Chapter 3

Chapter 3 is the methodology. This section comprises an Introduction, Type of Research Method, Research Design, Data Collection, Proposed Method, Research Tool, and Summary.

Chapter 4

The main chapter in this study is chapter 4 which consist of several steps to be done. This chapter will start from data preparation than shown the experimental and result discussion. Most of the experimental results are given in this chapter.

Chapter 5

Chapter 5 is conclusion and future work. This section presents the conclusion of this study. There is also describing about suggestion and recommendation for future work.

1.9 Chapter Summary

In conclusion, this chapter explains the background relate to direct marketing campaign. Therefore, the study to conduct precision prediction is required. Vary intelligent systems that are used to make decision in marketing campaign have been proposed, but these systems rarely for handle the imbalanced dataset.

Thus the objective of this study is to propose ANN for forecast or predict it is designed to handle the imbalance dataset. An improved ANN using K-Means algorithm for handling class imbalanced dataset seems to be a better solution to handle the particular problem.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discusses the foundation and literature relate to direct marketing in this literature review, the researcher investigates and explores the area that the reader may not know in relation to data mining techniques as well as previous applications. The reader can get a summarized view of the literature and analysis of the current issues involving the topic of the project. This would enable the reader to retrieve a wholesome summarized view of the topic.

2.2 Direct Marketing

Direct marketing is a marketing system that is interactive, utilizing the media to cause a measurable response or deals at any location (Chun, 2012). In direct marketing, promotional communications directed at individual consumers, with the aim that these messages addressed consumer is concerned, either by phone, mail or by coming directly to the consumer (Bose and Chen, 2009; Sing'oei and Wang, 2013). A model of influence campaign marketing forecast information on prices and advertising expenditures, and the expected benefits are administered by the quality of products, market acceptance of the product directly, and customer preferences (Sun, 2010).

2.3 Research Trend on Direct Marketing

Cui et al (2006) proposed a Bayesian Network learned by evolutionary programming with large direct marketing data set. They use data mining aided by advanced technology and versatile algorithm can remove many of the restrictions associated with traditional methods, and has become increasingly important as a new way to discover knowledge. They tested the endogeneity bias in the recency, frequency, monetary value (RFM) variables to control function by approach bayesian networks with those of neural networks, classification and regression tree (CART), and latent class regression. Furthermore they applied tenfold cross-validation with large database. They are adopt an innovative machine learning method of bayesian networks (BNS) learned by evolutionary programming (EP) to the model responses to direct marketing. They had realized the limitations of this method in the discrete nature. Although the discretization simplifies the learning process and the resulting model, there may be a loss of potentially useful information, and the model may not fully capture all the details of the relationship. EP can explore a wider search space to optimize BNS by comparing many alternatives, but this process may include valid models and affect the efficiency of the optimization process. Placing constraints on learning algorithms based on existing domain knowledge can help guide the search process to avoid any valid models and improve the overall efficiency and accuracy (Cui et al., 2006).

Crone et al (2006) concentrated study on the effect of different techniques of scale attributes prepossessing, sampling, coding categories and coding of continuous attributes in the classifier performance of decision trees, neural networks and support vector machines. They adopted multifactor direct marketing using analysis of variance on various performance metrics and methods parameterizations provide empirical evidence that the