CLASSIFICATION OF IMBALANCED DATASETS USING NAÏVE BAYES

NUR MAISARAH BINTI MOHD SOBRAN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical - Mechatronics & Automatic Control)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

MAY 2011

Dedicated to the one that believes in me,

“A small person with a big heart” -My mother

ACKNOWLEDGEMENT

Alhamdullillah, thanks to ALLAH swt for His bless at last I finished this project. First of all, I would like to take this opportunity to express my gratitude to my supervisor; Dr. Zuwairie Bin Ibrahim for encouragement supports, critics and helps. Without his guidance and interest, this project will not be a success.

I also would like to extend my appreciation to Asrul Adam who willing to help me to understand this project. With all his advice and guidance helps me a lot to complete this thesis.

Not to forget my beloved family, especially my parents for their fullest support throughout my two years of study in Universiti Teknologi Malaysia (UTM). It is because of them, I am the person who I am today.

I also would like to express my gratefulness to my employer, Universiti Teknikal Malaysia Melaka in providing me assistance in pursuing my master study. With their help I am able to concentrate in finishing this project.

My sincere appreciation also extends to all my fellow friends for their assistance and motivation at various occasions. Their views and tips are very useful indeed. Last but not least, thank you to all people who in one way or another contribute to the success of this project.

May Allah bless all of you.

Thank you.

ABSTRACT

Imbalanced data set had tendency to effect classifier performance in machine learning due to the greater influence given by majority data that overlooked the minority ones. But in classifying data, more important class is given by the minority data. In order to solve this problem, original Naïve Bayes was purposed as classifier for imbalanced data set. Our main interest is to investigate the performance of original Naïve Bayes classifier in imbalanced datasets. From the four UCI imbalanced datasets that been used, the purposed techniques show that, Naïve Bayes doing well in Herbaman's datasets and satisfying results in other datasets.

ABSTRAK

Ketidakseimbangan di dalam kumpulan data mempengaruhi kebolehan sistem mesin dalam mengelaskan data ke kelas masing-masing. Ini kerana "teknik pengelasan" yang digunakan dipengaruhi oleh kelas majoriti data walhal kelas data yang ingin dikenal pasti selalunya berada di kelas minoriti. Bagi mengatasi masalah ini, teknik pengelasan yang dipanggil "Naïve Bayes" telah digunakan terhadap kumpulan data yang tidak seimbang. Tujuan utama projek ini adalah untuk mengenalpasti tahap kebolehan Naïve Bayes dalam mengelaskan kumpulan data yang tidak seimbang. Hasil daripada pengaplikasian teknik ini terhadap empat kumpuan data, "Naïve Bayes" hanya menunjukkan keputusan yang baik terhadap kumpulan data Herbaman dan keputusan yang memberangsangkan terhadap kumpulan-kumpulan data yang lain.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

NB	-	Naïve Bayes
UCI	-	University of Carlifonia, Irvine
SVM	-	Support Vector Machine
SMOTE	-	Synthetic Minority Over Sampling Technique
NBS	-	Naïve Bayes Sampler
AESNB	-	Active Example Selection with Naïve Bayes Classifier
MLE	-	Maximum Likelihood Estimation
$P(X\|C)$	-	Probability joint distribution
$\prod$	-	Multipication
Σ	-	Summation
i	-	Number of attributes
n	-	Number of examples
j	-	Number of class
C	-	Class
X	-	Attributes
argmax	-	Maximum operator
N($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$)	-	Normal Distribution equation
μ	-	Mean
Σ	-	Covariance
θ_j	-	Parameter for j class
$\hat{\theta}$	-	Parameter estimation
D	-	Sampler with j class
$P(D,\theta)$	-	Parameter distribution from D sample
$\hat{\mu}$	-	Mean estimation
$\hat{\Sigma}$	-	Covariance estimation

ANN	-	Artificial Neural Network
kNN	-	k Nearest Neighbour
ABKN	-	Agent Based Knowledge Discovery
PSO	-	Particle Swam Optimization
GSVM-RU	-	Granular Support Vector Machine –Repetitive Under Sampling
CART	-	Classification and Regression tree
CNN	-	Complementary Neural Network
N-BR	-	Multi-relational Naïve Bayes

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of study

Adapting from human capability in learning from previous experience, researcher come out with methodologies for the machine to learn from prior dataset. The motivation, generally to come out with a computer system that improve from experience and capable in predicting the new outcome. In order to do that, a system must have capability in discovering knowledge in data set, model the pattern and from that model the system can predict future types of event that will happen.

One of common practice in machine learning is classification task. In classification, an algorithm was developed to discover knowledge from prior datasets. The prior datasets will provide information in terms of trends that available in each class. These trends will help the system in predicting the class for new instances. Because of this concept, a lot of attention had given towards classification method and eventually this concept largely implement in various fields. The applications cover medicine [1], industry [2], bussines and economy [3], fraud detection [4], remote sensing [5], and pattern recognition [6] area.

Due to the importance of classification task in machine learning, the best performance of learning algorithm is expected. So the practitioner made the algorithm with the objective of high accuracy in predicting the class and they assume the distribution between classes is the same or balanced. This assumption creates problem in real case implementation since there is difficulty in getting balanced dataset. With the imbalanced datasets problem, the classifier tends to favor the majority data (also called negative data) and treat the minority data (also called positive data) as noise but the important class is the minority class. For example, existing data for non cancer over cancer patient is 90% to 10%. If the classifier ignores the imbalanced data distribution, the results will be accuracy of 90% which is in majority data and ignore the important class of 10% cancer patients. But the important class to predict is on the minority data, cancer patients.

In this project, focus will be on studying the classification method that able to handling the imbalanced dataset. This project will give an insight view of implementation traditional classifier, Naïve Bayes. The contribution will be on the performance of traditional Naïve Bayes in imbalanced datasets and how much it differs from the other methods. This project will also discuss in what kind of environment that suitable in implementing this kind of method.

1.2 Problem statement

One of common classifier that been used for imbalanced datasets is Naïve Bayes [7]. However, in recent research activities on Naïve Bayes, the researchers tend to upgrade the original approach of Naïve Bayes method. It seems the original Naïve Bayes cannot uphold the best performance in classifying imbalanced datasets. Thus the researcher adds on sampling technique, features selection or mixture classifier agent with Naïve Bayes classifier in imbalanced dataset cases. But other problems arise. By implementing the Under Sampling will leads to the data loss in

classification task whereas the Oversampling method increase the simulation time and overlapped data may occurs. Other than that, mixtures of classifier agent add in complexity when implementing the method in real case scenario. Due to this problem, a study on performance of traditional Naïve Bayes Classification method is done in this project based on easy interpretation in modeling and algorithm. The comparison was made towards other methods in the end of this project.

1.3 Objective

The main objective of this project is to investigate the performance of original Naïve Bayes classification task for imbalanced datasets problems. It is hope that, in the end of project, this study will give initial overview towards original Naïve Bayes classifier performance.

1.4 Scope of study

This project will focus on the implementation of original Naïve Bayes in binary classification. Binary classification means either class “1” or class “0”. The datasets chosen consist of numbered value with multivariable inputs. This project is not dealing with any text classification datasets. The dataset been used based on the typical benchmarking datasets that usually implement for classification problem from University of California, Irvine (UCI) machine learning repository website without any data loss.

1.5 Outline of Project Report

This project was explained in five chapters. Chapter 1 states on the general idea of classification of imbalanced dataset, problem statement, the objective of project and scope of study.

Chapter 2 discover in more details the classification and imbalanced datasets concepts. After that overviews of previous approach done by researcher in the area will be discussed. Lastly, the writer will relates those literature with the one that used in this project, Naïve Bayes

Chapter 3 will describes the Naïve Bayes classification method and the implementation of this method in MATLAB programming environment.

Chapter 4 is for result and discussion section. In this chapter the performance of Naïve Bayes classification method was explained and comparison to other method was made.

Lastly, final conclusion was made in chapter 5. This chapter will conclude the result that we obtained in chapter 4 and come out with some suggestion for future work.

REFERENCES

1. Anuradha, B. and Reddy V.C.V.(2008). ANN for classification of Cardiac Arrhythmias. *ARPN Journal of Engineering and Applied Sciences*. Vol 3. pp1-5. Asian Research Publishing Newtwork.
2. Yip, W.K., Law, K.G., and Lee, W.J.(2007). Forecasting Final/Class Yeild on Fabrication Process E-Test and Sort Data. IEEE Conference on Automation, Science and Engineering. 22-25 Sept. Scottsdale,USA. pp 478-483.
3. Shihavuddin, A.S.M., Ambia, M.N., Arefin, M.M.N., Hossain, M. and Anwar, A. (2010). Prediction of stock price analyzing the online financial news using Naive Bayes classifier and local economic trends. *3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010*. 20-22 Aug. pp V4-22.
4. Padmaja, T.M., Dgulipalla, N., Bapi, S.R., and Krishna, P.R. (2007). Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. International Conference on Advanced Computing and Communications. 18-21 Dec. pp 511-516.
5. Bruzzone, L. and Serpico, S.B (1997). Classification of Imbalanced remote-sensing data by neural network. *Pattern Recognition Letters*,18, 1323-1328. Elsevier.
6. Porwal, A., Carranza, E.J.M, and Hale, M., (2006) Bayesian network classifier for mineral potential mapping. *Computing and Geosciences 32*. Elsevier Ptd. pp 1-16.
7. Derby, T. (2009). *Classification in imbalanced datasets*. Master of Science. Maastricht University.
8. Chawla, N.V., Japkowicz, N., and Kolcz, A. Editorial: Special Issue on learning from imbalanced datasets.
9. Provost. F. Machine Lerning from imbalanced datasets 101. Extended Abstract. New York University.
10. Kotsiantis, S.B., Kanellopoulos, D.,and Pintelas, P.E.(2006). Handling the imblalanced datasets: A review. GETS International Tarnsaction on Computer Science and Engineering. Vol 30.

11. Visa, S. *Fuzzy Classifier for Imbalanced datasets*. (2002).Doctor of philosophy. University of Cincinnati.
12. Sun, A., Lim, E., and Liu, Y. (2009). On Strategies of imbalnced text classification using SVM: A comparative study. Decision Support System 48. Pp 191-201. Elsevier Ptd.
13. Akbani, R.A, Kwek, S.A and Japkowicz, N.B.(2004). Applying support vector machines to imbalanced datasets. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)* .Volume 3201.39-50.
14. Tang, Y., Zhang, Y., Chawla, N.V. and Krasser, S.(2009) SVMs Modeling for Highly Imbalanced Classification. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics. Vol 39. pp281.
15. Dudo, R.O., Hart P.E., and Strok D.G.(2000) Pattern Classification 2nd edition. Wiley Interscience.
16. Quan, Z., Lin-gang, Gu., Chong-jun, W., Wang-jun and Shi-fu, Chen.,(2002). Using An Improved C4.5 for imbalanced datasets of intrusion. *AAMAS'02*. 15-19 July. Bologna, Italy.
17. Sug H.(2011). Improving the performance of Minor Class in Decision Tree Using Duplicating Instances. *AIKED'11 Proceedings of the 10th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases*.pp 234-237. World Scientific and Engineering Academy and Society (WSEAS).
18. Nguyen, G.H., Bouzerdoum, A., and Phong, S.L. (2008). A Supervised Learning Approach for Imbalanced Dataset. *International Conference on Pattern Recognition*. 8-11 Dec.
19. Yi Lu, Hong Guo and Feldkamp, L. (1998). Robust neural learning from unbalanced data samples. The 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence. 4-9 May. Vol 3. Pp 1816.
20. Jeatrakul, P., Wong, K.W., and Fung, C.C. (2010). Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE algorithm.*ICONIP 2010 (Part II).*pp 152-159. Springler-Verlag Berlin Heidelberg.

21. Zhong-Qiu Zhou. (2009). A novel modular neural network for imbalanced classification problem. Pattern Recognition Letters. Vol 30. pp 783-788. Elsevier B.V.
22. Soler, V., Cerquides ,J., Sabria, J., Roig, J., and Prim M. (2006). Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms. *ICDM Workshops 2006*. Hong Kong. 330.
23. Garcia, V., Mollineda, R.A., and Sanchez, J.S. (2007). On the k-NN performance in challenging scenario of imbalanced and overlapping. Pattern Anal Applic. Vol 11. pp 269-280. Springer-Verlag London Limited 2007.
24. Padmaja, T.M., Dhulipalla, N., Bapi, R.S. and Krishna, P.R. (2007). Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. *International Conference on Advanced Computing and Communications 2007*. 18-21 Dec. pp 511.
25. N.V.Chawla, L.O.Hall, K.W.Bowyer, and W.P.Kegelmeyer (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*. Vol.16, pp. 321-357.
26. Yun-chung A.(2004). *The effect of Oversampling and Undersampling on classifying Imbalanced Text Datasets*.Master of Science in Engineering. University of Texas, Austin.
27. Zhang, Q., Xue, Y., Zhou, H., and Tan, J. (2008). Research on Medical Documentation Categorization. *International Seminar on Future BioMedical Information Engineering 2008*. 18 Dec. pp 473.
28. Yang, P., Xu, L., Zhou, B.B., and Zomaya, A.T.(2009). A particle swam based hybrid system for imbalanced medical data sampling. *8th International Conference on Bioinformatics*. Singapore. 7-11 Sept. pp 1-14.
29. Lee, M.S., Rhee, J., Kim, Y., and Zhang B. (2009). AESNB: Active Example Selection with Naïve Bayes Classifier for Learning from Imbalanced Biomedical Data. *Ninth IEEE International Conference on Bioinformatics and Bio Engineering, 2009*. BIBE '09. 22-24 June. pp 15.
30. Xu, G., Bao, H., and Meng, X.(2008). Multi-Relational Classification in Imbalanced Domains. Lecture Notes in Computer Science. *Advances In Computation And Intelligence 2008*. Volume 5370/2008. pp 562-570. Springer-Verlag Berlin Heidelberg 2008.

31. Wijayatunga, W.J..P.S.P. (2007). *Statistic Analysis and Application of Naïve Bayesian Network*. Classsifier.Doctor of Philosophy. Tokyo Institude of Technology.
32. Needham, C.J., Bradford, J.R., Bulpitt, A.J., and Westhead, D.R. (2007). A Primier on learning in Bayesian Networks for Computational Biology. Vol 3. pp 129. PLoS Computational Biology.
33. Mitchell, T.M. (2010). *Machine Learning*. 2nd edition draft. McGraw Hill.
34. Anderson, T.W. (2003).An Introduction to Statistical Analysis. 3rd Edition. United State of America.A John Wiley & Sons .Inc.
35. Gu, Q., Zhu, L., and Cai, Z., (2009). Evaluation Measures of the Classification Performance of the Imbalanced datasets. *ISICA 2009*. Pp 461-471. Spriger-Verlag Berlin Heidelberg 2009.
36. Kotsiantis, S.B., and Pintelas, P.E.(2003). Mixture of Expert Agents for Handling Imbalanced Datasets. Annals of Mathematic, Computing, & Teleinformatics. Vol 1. pg 46-55.
37. Giang H. Nguyen, Abdesselam Bouzerdoum, and Son L. Phung, (2008) "A Supervised Learning Approach for Imbalanced Data Sets", Proceeding of the19th International Conference on Pattern Recognition (ICPR 2008), pp. 1-4.
38. Phung, S.L., Bouzerdoum, A., and Nguyen, G.H. (2009). Learning pattern classification task with imbalanced datasets. 193-208. Research Online. University of Wollongong.