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.