Complexity Measurement for Dealing with Class Imbalance Problems in Classification Modelling

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

Massey University, 2012

Muhammad Nafees Anwar

Institute of Fundamental Sciences

Massey University

New Zealand

To my beloved late Mother

Abstract

The class imbalance problem is a challenge in the statistical, machine learning and data mining domains. Examples include fraud/intrusion detection, medical diagnosis/monitoring, bioinformatics, text categorization, insurance claims, and target marketing. The problem with imbalanced data sets is that the conventional classifiers (both statistical and machine learning algorithms) aim at maximizing overall accuracy, which is often achieved by allocating all, or almost all, cases to the majority class. Thus there tends to be bias against the minority class in class imbalance situations.

Despite numerous algorithms and re-sampling techniques proposed in the last few decades to tackle imbalanced classification problems, there is no consistent winning strategy for all data sets (neither in terms of sampling, nor learning algorithm). Special attention needs to be paid to the data in hand. In doing so, one should take into account several factors simultaneously: the imbalance rate, the data complexity, the algorithms and their associated parameters. As suggested in the literature, mining such datasets can only be improved by algorithms tailored to data characteristics; therefore it is important and necessary to do data exploratory analysis before deciding on a learning algorithm or re-sampling techniques.

In this study, we have developed a framework "Complexity Measurement" (CM) to explore the connection between the imbalanced data problem and data complexity. Our study shows that CM is an ideal candidate to be recognized as a "goodness criterion" for various classifiers, re-sampling and feature selection techniques in the class imbalance framework. We have used CM as a meta-learner to choose the classifier and under-sampling strategy that best fits the situation. We design a systematic over-sampling technique, Over-sampling using Complexity Measurement (OSCM) for dealing with class overlap. Using OSCM, we do not need to search for an optimal class distribution in order to get favorable accuracy for the minority class, since the amount of over-sampling is determined by the complexity; ideally using CM would detect fine structural differences (class-overlap and small disjunct) between different classes.

Existing feature selection techniques were never meant for class imbalanced data. We propose Feature Selection using Complexity Measurement (FSCM), which can specifically focus on the minority class, hence those features (and multivariate interactions between predictors) can be selected, which form a better model for the minority class.

Methods developed have been applied to real datasets. The results from imbalanced datasets show that CM, OSCM and FSCM are effective as a systematic way of correcting class imbalance/overlap and improving classifier performance. Highly predictive models were built; discriminating patterns were discovered, and automated optimization was proposed. The methodology proposed and knowledge discovered will benefit exploratory data analysis for imbalanced datasets. It may be taken as a judging criterion for new algorithms and re-sampling techniques. Moreover, new data sets may be evaluated using our CM criterion in order to build a sensible model.

Acknowledgements

I am highly in debt to Dr. Geoff Jones for his invaluable guidance and help to me in the development of my research skills, and his efforts to keep a stray bird on the right track.

I am very grateful to Dr. Siva Ganesh. Ganesh has always allowed me complete freedom to define and explore my own directions in research. While this proved difficult and somewhat bewildering to begin with, I have come to appreciate the wisdom of his way, as it encouraged me to think for myself.

I am thankful to Dr. Martin Hazelton and the Statistic Group who have provided me with financial support during my degree. They have kindly provided casual teaching assistantship positions and travel funds to attend conferences.

I thank Dr. Ganes Ganesalingam and Dr. Mark Bebbington for providing valuable feedback on parts of this thesis.

I also thank Mr. Peter Lewis for providing technical support.

Special thanks must also go to my family, especially to my late mother and to my wife Wa. They have provided unconditional support and encouragement through both the highs and lows of my time in Ph.D study.

Finally I would like to thank Higher Education Commission (HEC), Pakistan, for providing me with a scholarship to study at Massey University, New Zealand.

Contents

List of Figures

List of Tables