# Multiple Costs and Their Combination in Cost Sensitive Learning

By

Zhenxing Qin

Submitted in fulfilment of the requirement for the degree of

**Doctor of Philosophy**

**University of Technology, Sydney**

June 2006

# CERTIFICATE OF AUTHORSHIP/ORIGINALITY

*I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.*

*I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.*

*Signature of Candidate*

Production Note:
Signature removed prior to publication.

*To my wife Yanfang and Our Parents*

# Abstract

Cost sensitive learning is firstly defined as a procedure of minimizing the costs of classification errors. It has attracted much attention in the last few years. Being cost sensitive has the strength to handle the unbalance on the misclassification errors in some real world applications. Recently, researchers have considered how to deal with two or more costs in a model, such as involving both of the *misclassification costs* (the cost for misclassification errors) and *attribute test costs* (the cost incurs as obtaining the attribute's value) [Tur95, GGR02, LYWZ04]. Cost sensitive learning involving both attribute test costs and misclassification costs is called ***test cost sensitive learning*** that is more close to real industry focus, such as medical research and business decision.

Current test cost sensitive learning aims to find an optimal diagnostic policy (simply, a policy) with minimal expected sum of the misclassification cost and test cost that specifies, for example which attribute test is performed in next step based on the outcomes of previous attribute tests, and when the algorithm stops (by choosing to classify). A diagnostic policy takes the form of a decision tree whose nodes specify tests and whose leaves specify classification actions. A challenging issue is the choice of a reasonable one from all possible policies.

This dissertation argues for considering both of the test cost and misclassification cost, or even more costs together, but doubts if the current way, summing up the two costs, is the only right way. Detailed studies are needed to ensure the ways of combination make sense and be "correct", dimensionally as well as semantically. This

dissertation studies fundamental properties of costs involved and designs new models to combine the costs together.

Some essential properties of attribute test cost are studied. In our learning problem definition, test cost is combined into misclassification cost by choosing and performing proper tests for a better decision. Why do you choose them and how about the ones that are not chosen? Very often, only part of all attribute values are enough for making a decision and rest attributes are left as "unknown". The values are defined as '*absent values*' as they are left as unknown purposely for some rational reasons when the information obtained is considered as enough, or when patients have no money enough to perform further tests, and so on.. This is the first work to utilize the information hidden in those "absent values" in cost sensitive learning; and the conclusion is very positive, i.e. "Absent data" is useful for decision making. The "absent values" are usually treated as '*missing values*' when left as known for unexpected reasons. This thesis studies the difference between 'absent' and 'missing'. An algorithm based on lazy decision tree is proposed to identify the absent data from missing data, and a novel strategy is proposed to help patch the "real" missing values. .

Two novel test cost sensitive models are designed for different real work scenarios. The first model is a *general test cost sensitive learning framework with multiple cost scales*. Previous works assume that the test cost and the misclassification cost must be defined on the same cost scale, such as the dollar cost incurred in a medical diagnosis. And they aim to minimize the sum of the misclassification cost and the test cost. However, costs may be measured in very different units and we may meet difficulty in defining the multiple costs on the same cost scale. It is not only a technology issue, but

also a social issue. In medical diagnosis, how much money should you assign for a misclassification cost? Sometimes, a misclassification may hurt a patient's life. And from a social point of view, life is invaluable. To tackle this issue, a *target-resource budget learning framework* with multiple costs is proposed. With this framework, we present a test cost sensitive decision tree model with two kinds of cost scales. The task is to minimize one cost scale, called target cost, and keep the other one within specified budgets. To the best of our knowledge, this is the first attempt to study the cost sensitive learning with multiple costs scales.

The second model is based on the assumption that some attributes of an unlabeled example are known before being classified. A test cost sensitive lazy tree model is proposed to utilize the known information to reduce the overall cost. We also modify and apply this model to the batch-test problem: multiple tests are chosen and done in one shot, rather than in a sequential manner in the test-sensitive tree. It is significant in some diagnosis applications that require a decision to be made as soon as possible, such as emergency treatment.

Extensive experiments are conducted for evaluating the proposed approaches, and demonstrate that the work in this dissertation is efficient and useful for many diagnostic tasks involving target cost minimization and resource utilization for obtaining missing information.

# Acknowledgements

First, I would like to take this opportunity to express my sincere gratitude to my supervisor, Professor Chengqi Zhang, for his unreserved encouragement, advice and support, and for giving me the opportunity to pursue my PhD at the Faculty of Information Technology at the University of Technology, Sydney. In particular, when I did not meet the enrolment deadline in 2003 due to the IELTS and Visa delay, he made much effort to hold my scholarship for more than half a year until I had the honour of studying and working with him in the past three years, and that is stamped indelibly on my life. His comments are so helpful, and his suggestions are so challenging, that I profit a lot from his guidance during my PhD program. I am honoured and feel happy to be able to work under such a supervisor like Chengqi. His profound knowledge and wisdom have deeply impressed me. I will remember his kindness forever.

Also, I am full of gratitude to Dr. Shichao Zhang, my co-supervisor, for his great help on my daily life and study, and for his detailed and constructive advice on my researches. He is also my master supervisor in China and he recommends me to Professor Chengqi Zhang as his Ph. D student, i.e. he is always a pilot of my life road and guides my onward direction. I would not be able to complete this thesis without his selfless help. He has always provided me with his knowledgeable views.

I am grateful to the Faculty of Information Technology, the University of Technology, Sydney, for providing me with a nice opportunity, an excellent environment and scholarship for the learning and researching here. I am grateful to Dr. Jie Lu and Dr. Guangquan Zhang for their kindl support and valuable discussion.

My thanks also go to all those who have helped me in one way or another during my PhD course: Ms. Li Liu, Ms Yanchun Zhou, Mr. Xiaowei Yan, Mr. Longbing Cao, Mr. Qingfeng Cheng, Mr. Chunsheng Li, Mr. Li Lin, Mr. Jiaqi Wang, Mr. Wanli Chen, Mr. Jiarui Ni, Mr. Yanchang Zhao, Mr. Jiarui Ni, Mr. Chengeng Shi, Mr. Xuetao Guo, and Mr. Alan Tao Wang.

Thanks to my wife Yanfang Ji, and our parents. They give me a happy feeling of family.

In addition, I wish to thank the University of California, Department of Information and Computer Science for providing the UCI Repository of machine learning databases [BM98].

# List of Publications

The following is a list of my research papers published in the proceedings of referred international conferences or journals during my PhD study at University of Technology, Sydney.

**Referred Journal Papers:**

1 Shichao Zhang, Zhenxing Qin, Charles Ling and Shengli Sheng, "Missing is Useful": Missing Values in Cost-sensitive Decision Trees," **IEEE Transactions on Knowledge and Data Engineering**, Vol. 17 No. 12 (2005): 1689-1693.

2 Chengqi Zhang, Zhenxing Qin, Xiaowei Yan, "Association-Based Segmentation for Chinese-Crossed Query Expansion," **IEEE Intelligent Informatics Bulletin** 5(1), 2005: 18-25.

**Referred Conference Papers:**

3 Zhenxing Qin, Chengqi Zhang and Shichao Zhang, "Missing or absent? A Question in Cost-sensitive Decision Tree," **Proceedings of the Fourth International Conference on Active Media Technology** (AMT006), Jun 2006:

4 Yiming Yang, Qiang Yang, Rong Pan et al. and Zhenxing Qin, "Preprocessing Time Series Data for Classification with Application to CRM," In **Proceedings of the 18th Australian Joint Conference on Artificial Intelligence** (AI 2005), Sydney, Australia, 2005: 133-142.

5 Zhenxing Qin, Chengqi Zhang and Shichao Zhang, "Dynamic Test-sensitive Decision Trees with Multiple Cost Scales," In **Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery** (FSKD-2005), Changsha, China, August 2005: 402-405.

6 Zhenxing Qin, Chengqi Zhang and Shichao Zhang, "Cost-sensitive Decision Trees with Multiple Cost Scales," In **Proceedings of the 17th Australian Joint Conference on Artificial Intelligence** (AI 2004), Cairns, Queensland, Australia, 2004: 380-390.

7 Zhenxing Qin, Li Liu and Shichao Zhang, "Mining Term Association Rules for Heuristic Query Construction," In **Proceedings of 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining**, (PAKDD 2004) Sydney, Australia, May 26-28, 2004: 145-154.

# Table of Contents

# List of Figures