Faculty of Engineering and Information Technology
University of Technology, Sydney

Ensemble Predictions: Empirical Studies on Learners' Performance and Sample Distributions

A thesis submitted in fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Guohua Liang

February 2014

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

Acknowledgments

I would like to express my deepest gratitude to my principal supervisor Professor Chengqi Zhang for providing me with such an excellent research environment; for the wonderful opportunities to explore my dreams to become a successful scientist; for his supervision; ongoing encouragement and support. His deep insights, profound knowledge and outstanding contributions to our society have been a great model for me. All of these will benefit my future research career. It has been a great honor for me to be his PhD student.

I would especially like to express my deepest appreciation to my co-supervisor Professor Xingquan Zhu. His guidance and rigorous approach in all academic areas has been of great value, allowing me to make significant academic achievements in a short period of time. I am very grateful for his support and advice.

My warm thanks to all members, visitors and students of the Centre for Quantum Computation & Intelligent Systems, University of Technology, Sydney for their helpful discussions and encouragement.

I would like to dedicate this thesis to my family, my husband Bin Kong and my two sons, Jeffrey Kong and Kevin Kong for their love, support, and understanding, especially, my little son, Kevin. He always asks me, “Mum why you always go to your office doing your work?” Sometimes he talks to himself, “My mum has a lot of work to do”; he also frequently says to me, “When you finish your work come back home, I will give you some gold stars.”

Finally, I would like to express my gratitude to both the Graduate School and the Faculty of Engineering and IT, University of Technology, Sydney for offering financial support with the final stages of my thesis writing.

Contents

Certificate . . . . **i**

Acknowledgment . . . . **ii**

Abstract . . . . **xvi**

Chapter 1 Introduction **1**

1.1 Objectives 3

1.2 Contributions 5

1.3 Organisation 7

1.4 Publications Related to the Thesis 7

Chapter 2 Related Work **10**

2.1 Ensemble Learning 10

2.1.1 General Ensemble Learning Approaches 11

2.1.2 Empirical Studies on Ensemble Learning 13

2.2 Bagging 14

2.2.1 Basic Concept and Framework of Bagging 14

2.2.2 Bagging Algorithm 15

2.2.3 Advantages of Bagging 16

2.2.4 Bagging Background and Approaches 16

2.3 Statistical Test 19

2.3.1 Wilcoxon Signed-rank Test 19

2.3.2 Friedman Test and Post-hoc Nemenyi Test 20

2.4 Sampling Techniques 21
2.5 Evaluation Metrics 24
2.5.1 *ROC* 27
2.5.2 How to Calculate *AUC* of *ROC* 28
2.6 Basic Learning Algorithms 29
2.7 Benchmark Data-sets 30
2.7.1 Imbalanced Data-sets 30
2.7.2 Selection of Medical Data-sets 31

Chapter 3 An Effective Approach for Imbalanced Classification: UBagging 34
3.1 Introduction 35
3.2 The UBagging Algorithm 38
3.3 Related Work 40
3.4 Experimental Setup 41
3.4.1 Data-sets 43
3.5 Experimental Results and Analysis 43
3.6 Conclusion 48

Chapter 4 An Empirical Study of Bagging Predictors with Different Learning Algorithms 49
4.1 Introduction 50
4.2 Designed Framework 52
4.3 Base Learner Characterization 53
4.4 Experimental Setting 55
4.5 Experimental Analysis 56
4.5.1 Comparison of All Bagging Predictors 57
4.5.2 Comparison of Two Learners Bagging and Single Learner 60
4.5.3 Comparison of Average Improvement of Bagging 60
4.6 Conclusions 62

Chapter 5 An Empirical Study of the Sensitivity of Bagging on Imbalanced Class Distribution 63
5.1 Introduction 64
5.2 Designed Framework 66
5.2.1 Sensitivity of Bagging Predictor 67
5.2.2 Friedman Test with Post-hoc Nemenyi Test 68
5.2.3 Evaluation Metrics 69
5.3 Experimental Results 69
5.3.1 Statistical Analysis 69
5.3.2 Graphical Analysis 73
5.4 Conclusion 74

Chapter 6 The Impact of Class Distribution on Bagging . . 75
6.1 Introduction 76
6.2 Designed Framework 79
6.2.1 Random Under-sampling Technique Varying the Levels of Class Distribution 81
6.3 Experimental Setting 82
6.4 Experimental Results Analysis 83
6.4.1 Statistical Comparison Bagging Predictors with Single Learners 84
6.4.2 Graphical Comparison of *ROC* Curves 87
6.4.3 Comparison of the Performance of All Bagging Predictors 90
6.5 Conclusion 97

Chapter 7 An Empirical Investigation of Bagging on Domain Specific Data 99
7.1 Introduction 100
7.2 Designed Framework 104
7.3 Experimental Setting 104

7.4 Experimental Results Analysis 105

7.4.1 Comparison of Bagging with Single Learners 105

7.4.2 Comparison of All Bagging Predictors 106

7.4.3 Comparison of the Performance of Prediction Models on Individual Medical Data-sets 110

7.4.4 Comparison of the Performance of Bagging between Natural Class Distribution and Altered Class Distribution on Individual Medical Data-sets 117

7.5 Conclusions 122

Chapter 8 An Effective Method for Imbalanced Time Series Classification: Hybrid Sampling 124

8.1 Introduction 126

8.2 HBagging approach 127

8.2.1 Statistical Tests 129

8.3 Experimental Setup 130

8.3.1 Data-sets 130

8.4 Experimental Results Analysis 131

8.4.1 Evaluation of the Performance of SVM 132

8.4.2 Comparison of Over-sampling, Under-sampling, and Hybrid-sampling Methods 133

8.4.3 Comparison of the Performance of State-of-the-art Methods in TSC, SPO, Under-sampling, and H-sampling Methods 136

8.5 Conclusion 139

Chapter 9 Conclusions and Future Work 140

9.1 Conclusions 140

9.2 Future Work 143

Bibliography **144**

List of Figures

2.1 Framework of bagging 14

3.1 Comparison of average rank of F_{value} of the performance of four prediction models with the Nemenyi test, where the x-axis indicates the average rank of F_{value}, the y-axis indicates the ranking order of the four prediction models, and the vertical bars indicate the "critical difference". 47

3.2 Comparison of average rank of G_{mean} of the performance of four prediction models with the Nemenyi test, where the x-axis indicates the average rank of G_{mean}, the y-axis indicates the ranking order of the four prediction models, and the horizontal bars indicate the "critical difference". 47

4.1 Designed framework 53

4.2 Two-dimensional robustness and stability decomposition of the base learners based on estimated error rate and variance, where the x-axis denotes the robustness of the base learners from robust to weak, and the y-axis denotes the stability of the base learners from stable to unstable. 54

4.3 Friedman and post-hoc Nemenyi test results of comparison of all bagging predictors, where the x-axis indicates the mean rank of bagging predictors, the y-axis indicates the ranking order of the bagging predictors, and the horizontal error bars indicate the "critical difference". 59

4.4 The improved accuracy between bagging predictors and individual base learners on average over multiple data-sets. The error bars present a 95% confidence interval based on the cross-validated t-test. 61

5.1 Designed framework 67

5.2 Comparison of all bagging predictors with the Nemenyi test, where the x-axis indicates the average rank of the bagging predictors, the y-axis indicates the ascending order of the average rank of CG performance, and the horizontal bars indicate the CD. 71

5.3 Comparison of ROC curve and G_{mean} among selected bagging predictors and data-sets. 72

6.1 Designed framework 79

6.2 Comparisons of ROC curves between a B_MLP and a single learner MLP on 12 imbalanced data-sets, where the x-axis denotes FPR, the y-axis denotes TPR for each sub-figure. . 88

6.3 The group of comparisons of ROC curves between 12 bagging predictors and single learners on the *Diabetes* data-set, where the x-axis denotes FPR, the y-axis denotes TPR for each sub-figure. 89

6.4 Comparison of the performance of all bagging predictors with post-hoc Nemenyi test, where x-axes indicate the mean rank of G_{mean} for bagging, the y-axes indicate the ascending ranking order of the bagging predictors and the horizontal error bars indicate the "critical difference". 92

6.5 Comparison of the TPR performance of all bagging predictors with the Nemenyi test, where the x-axes indicate the mean rank of TPR for bagging predictors, the y-axes indicate the ascending ranking order of the bagging predictors, and the horizontal bars indicate the "critical difference". 94

6.6 Average ranks of AUC performance for 12 bagging predictors with the Nemenyi test, where the x-axis denotes the ranking order of the bagging predictors, while the y-axis denotes the average rank of the AUC performance of the bagging predictors. The error bars present the "critical difference" of the Nemenyi test. 94

6.7 The average AUC performance of bagging predictors over 14 data-sets, where the x-axis indicates the name of the bagging predictors, the y-axis indicates the average value of AUC and the error bar indicates the variance value. 96

7.1 Designed framework 103

7.2 Comparison of the G_{mean} performance of all bagging predictors with post-hoc Nemenyi test, where the x-axes indicate the mean rank of each bagging predictor, the y-axes indicate the ascending ranking order of the bagging predictors, and the vertical error bars indicate the "critical difference". 109

7.3 The performance of prediction models on *Breastc* data-set. . . 110

7.4 Comparison of the performance of prediction models on *Diabetes* data-set. 111

7.5 Comparison of the performance of prediction models on *Sick* data-set. 111

7.6 Comparison of the performance of bagging predictors and single learners on *Heart-h* data-set. 111

7.7 Comparison of the performance of the bagging predictors and single learners on $WDBC$ data-set. 113

7.8 Comparison of the performance of the bagging predictors and single learners on *Heart-c* data-set. 113

7.9 Comparison of the performance of the bagging predictors and single learners on $WBreastc$ data-set. 113

7.10 Comparison of the performance of the bagging predictors and single learners on $StatlogHeart$ data-set. 114

8.1 Comparison of average rank of the F_{value} with the Nemenyi test for the over-sampling methods, under-sampling with with various algorithms, and HBagging, where the x-axis indicates the ranking order of the average rank of the F_{value}, the y-axis indicates the average rank of the F_{value}, and the vertical bars indicate the "critical difference". 134

8.2 Comparison of average rank of the G_{mean} with the Nemenyi test for all the over-sampling methods, under-sampling with various algorithms, and HBagging method, where the x-axis indicates the ranking order of the average rank of the G_{mean}, the y-axis indicates the average rank of the G_{mean}, and the vertical bars indicate the "critical difference". 135

8.3 Comparison of average rank of the F_{value} metric with the Nemenyi test for the state-of-the-art methods in TSC, SPO, under-sampling with various algorithms, and HBagging, where the x-axis indicates the ranking order of the average rank of F_{value}, the y-axis indicates the average rank of F_{value}, and the vertical bars indicate the "critical difference". 138

8.4 Comparison of average rank of the G_{mean} metric with the Nemenyi test for the state-of-the-art methods in TSC, SPO, under-sampling with various algorithms, and HBagging, where the x-axis indicates the ranking order of the average rank of G_{mean}, the y-axis indicates the average rank of G_{mean}, and the vertical bars indicate the "critical difference". . 138

List of Tables

2.1 Confusion matrix for a binary classification problem 25

2.2 Experimental data-set characteristics 32

2.3 Imbalanced data-sets 33

2.4 Medical data-sets 33

3.1 Imbalanced data-sets (ordered by $\%P$) 42

3.2 Comparison of the performance of four prediction models based on F_{value} and G_{mean} 44

4.1 Mean rank of Friedman test for error rate of bagging predictors 57

4.2 The results of the Wilcoxon signed-rank test to compare the estimated $ErrorRate$ of bagging and single learners. The significance level is .05. 59

5.1 Statistical results of Wilcoxon signed-rank test 69

6.1 Under-sampling technique altering the levels of class distribution 81

6.2 Sampling techniques are used to change each original data-set into 9 altered data-sets with 9 levels of class distribution for building 9 single and bagging final prediction models, respectively. These prediction models produce 9 pairs (FPR, TPR) to form a *ROC* curve for single and bagging prediction models, respectively. 83

6.3 The statistical results of the Wilcoxon signed-rank test for comparison of the G_{mean} performance of bagging and single learners. The significance level is .05. 84

6.4 The statistical results of the Wilcoxon signed-rank test for comparison of the TPR performance of bagging and single learners. The significance level is .05. 86

6.5 The statistical results of the Wilcoxon signed-rank test for comparison of the AUC performance of bagging and single learners. The significance level is .05 86

6.6 Mean rank of the Friedman test for G_{mean} performance of bagging predictors 91

6.7 Mean rank of the Friedman test for TPR performance of bagging predictors 92

6.8 Mean rank of the Friedman test for AUC performance of bagging predictors 93

6.9 Average AUC performance of bagging predictors on 14 imbalanced data-sets 96

7.1 Compare bagging with each single learner based on Wilcoxon signed-rank test on G_{mean}. The significance level is .05. 106

7.2 Ranking order of the performance of bagging based on G_{mean} and mean ranks. 108

7.3 Best G_{mean} performance prediction models for the natural class distribution on individual medical data-sets 115

7.4 The best G_{mean} performance of the bagging prediction models achieved with altered class distribution on individual data-sets 116

7.5 Comparison of the performance of bagging predictors on *Breastc* and *Heart-c* data-sets 118

7.6 Comparison of the performance of bagging predictors on *Heart-h* and *StatlogHeart* data-sets 119

7.7 Comparison of the performance of bagging predictors on *Diabetes* and *Sick* data-sets 120

7.8 Comparison of the performance of bagging predictors on *WDBC* and *WBreastc* data-sets 121

8.1 Time series data-sets 130

8.2 Results of SVM on imbalanced time series data-sets 132

8.3 Comparison of the performance of over-sampling methods, under-sampling with various algorithms, and HBagging method based on the evaluation metrics F_{value} and G_{mean}. . . 133

8.4 Comparison of the performance of state-of-the-art methods in TSC, SPO, under-sampling with various algorithms, and HBagging based on evaluation metrics: F_{value} and G_{mean}. . . . 137

Abstract

Imbalanced data problems are among the most challenging in Data Mining and Machine Learning research. This dissertation investigates the performance of ensemble learning systems on different types of data environments, and proposes novel ensemble learning approaches for solving imbalanced data problems. Bagging is one of the most effective ensemble methods for classification tasks. Despite the popularity of bagging in many real-world applications, there is a major drawback on extremely imbalanced data. Much research has addressed the problems of imbalanced data by using over-sampling and/or under-sampling methods to generate an equally balanced training set to improve the performance of the prediction models. However, it is unclear which is the best ratio for training, and under which conditions bagging is outperformed by other sampling schemes on extremely imbalanced data.

Previous research has mainly been concerned with studying unstable learners as the key to ensuring the performance gain of a bagging predictor, with many key factors remaining unclear. Some questions have not been well answered: (1) What are the key factors for bagging predictors to achieve the best predictive performance for applications? and (2) What is the impact of varying the levels of class distribution on bagging predictors on different data environments. There is a lack of empirical investigation of these issues in the literature.

The main contributions of this dissertation are as follows:

1. This dissertation proposes novel approaches, uneven balanced bagging to boost the performance of the prediction model for solving imbalanced problems, and hybrid-sampling to enhance bagging for solving highly imbalanced time series classification problems.

2. This dissertation asserts that robustness and stability are two key factors for building a high performance bagging predictor. This dissertation also derives a new method, utilizing two-dimensional robustness and stability decomposition to rank the base learners into different categories for the purpose of comparing the performance of bagging predictors with respect to different learning algorithms. The experimental results demonstrate that bagging is influenced by the combination of robustness and instability, and indicate that robustness is important for bagging to achieve a highly accurate prediction model.

3. This dissertation investigates the sensitivity of bagging predictors. We demonstrate that bagging MLP and NB are insensitive to different levels of imbalanced class distribution.

4. This dissertation investigates the impact of varying levels of class distribution on bagging predictors with different learning algorithms on a range of data environments, to allow data mining practitioners to choose the best learners and understand what to expect when using bagging predictors.