



Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand). This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library

[Thesis Consent Form](#)

Ensemble Learning by Data Resampling

Michael Goebel

A thesis submitted in partial fulfilment of the requirements for
the degree of Doctor of Philosophy in Computer Science,
The University of Auckland, 2004

Abstract

We investigate ensemble learning methods that construct a classifier ensemble by repeatedly sampling the original training data and building a member classifier from each subsample. We find that the performance of standard Bagging can frequently be improved upon by simple variations of the sampling scheme, such as varying the sample size, or sampling without replacement instead of sampling with replacement.

For all methods tested, the ensemble performance is greatly dependent on properties of the problem domain and the data sample. We try to explain the observed performances of the various ensemble methods qualitatively and quantitatively, and find that current ensemble analysis methods such as margin distributions, κ -Error Diagrams, bias-variance decomposition etc. are not well suited for this task.

We postulate that the primary explanation for the performance of ensemble methods is to be found in their effects on accuracy of the ensemble members on one side and diversity among them on the other side, two contradictory goals which necessitate a compromise, or trade-off. This motivates the presentation of a precise yet general definition of what diversity is and how it is to be measured. This definition has the desirable property that it is applicable to all single-stage voting ensembles, under any given loss function.

We then study the mathematical relationships between ensemble loss, mean member loss, and diversity. For squared loss, we show that our definitions lead to the well known ensemble loss decomposition, and extend this decomposition to the case where the ensemble members, instead of a real number, return a probability distribution over \mathcal{R} .

For the case of 0-1 loss, we derive the exact mathematical relations between ensemble loss, mean member loss, and diversity. Studying those relations provides some valuable insights into ensembles behavior, and produces some unexpected hence interesting results. These results are also confirmed by the experimental observations.

Turning our attention back to the performance of Bagging variants, we show how the loss decomposition can be used to reduce the number of parameter settings which have to be tried out experimentally in order to find a well-performing ensemble method for a given particular problem.

Acknowledgments

I am greatly indebted to my supervisory committee, Pat Riddle, Mike Barley, and Hans Guesgen, for innumerable helpful comments, as well as for their general guidance and support.

Another big ‘Thank you’ to all those who provided valuable comments on earlier versions of this thesis, especially to Remco Bouckaert.

Special thanks also goes to the Department of Computer Science at the University of Auckland for their financial, technical, and administrative support.

Appendix A lists those who provided the datasets used in the experiments.

Credits are also due to all those who contribute their time and energy to produce all these wonderful free (‘free’ as in free speech) software packages used for the conduct of the research as well as for the production of this thesis – you are way too many to mention individually, but way too important not to mention at all.

Lastly, to my parents, Rotraut and Manfred, and to my fiancée, María Cristina, as well as to her parents, María del Carmen and José Carlos, for their extraordinary patience, love, and support.

This would not have been possible without you.

Contents

| | |
|--|-------------|
| List of Tables | vi |
| List of Figures | viii |
| Abbreviations | xi |
| Notation | xii |
| 1 Introduction | 1 |
| 1.1 Scope | 1 |
| 1.2 Motivation | 1 |
| 1.3 Scientific Contributions | 3 |
| 1.4 Organization | 4 |
| 2 Problem Description | 6 |
| 2.1 Classification and Regression | 6 |
| 2.2 Loss Functions | 7 |
| 2.3 Ensembles | 10 |
| 2.4 Bagging and Cragging | 14 |
| 2.5 Relationship to Bayesian Model Averaging | 16 |
| 3 The Accuracy-Diversity Trade-Off | 18 |
| 4 Current Ensemble Analysis Methods | 24 |
| 4.1 Experimental Methodology | 24 |
| 4.2 Error Curves | 26 |
| 4.3 κ -Error Diagrams | 38 |

| | | |
|----------|--|-----------|
| 4.4 | Margin Distributions | 42 |
| 4.5 | Bias-Variance Decomposition | 45 |
| 5 | Loss Decomposition | 50 |
| 5.1 | Voting Schemes | 50 |
| 5.2 | Decomposition | 53 |
| 5.3 | Instantiating the Decomposition for Squared Loss | 55 |
| 5.4 | Instantiating the Decomposition for 0-1 Loss | 56 |
| 5.4.1 | 0-1 Loss for Two-Class Problems | 56 |
| 5.4.2 | 0-1 Loss for Multi-Class Problems | 58 |
| 5.4.3 | Discussion | 59 |
| 5.5 | General Bounds on the Expected Ensemble Loss | 63 |
| 6 | Applying the Loss Decomposition | 64 |
| 6.1 | Probabilistic vs. Majority Vote | 66 |
| 6.2 | Comparison of Sampling Schemes | 67 |
| 6.3 | Number of Classes | 75 |
| 7 | Conclusions and Further Research | 77 |
| 7.1 | Summary | 77 |
| 7.2 | Further Research | 78 |
| 7.2.1 | Weighted Ensembles | 78 |
| 7.2.2 | Other Ensemble Learning Methods | 79 |
| 7.2.3 | Other Loss Functions | 80 |
| 7.2.4 | Other Voting Functions | 80 |
| 7.2.5 | Other Base Learners | 80 |
| 7.2.6 | Multi-Stage Ensembles | 81 |
| 7.3 | Conclusions | 82 |
| A | Dataset Providers | 83 |
| B | Performance of Base Classifier | 85 |
| C | Error curves for $Cragging(n; 1)$ | 86 |

| | | |
|----------|---|------------|
| D | κ-Error Diagrams | 88 |
| E | Cumulative Margin Distributions | 103 |
| F | Bias-Variance Decomposition Results | 124 |
| G | Proofs | 137 |
| | G.1 Proof of Theorem 5.1 | 137 |
| | G.2 Proof of Theorem 5.2 | 139 |
| | G.3 Proof of Theorem 5.3 | 140 |
| | G.4 Proof of Theorem 5.4 | 141 |
| | G.5 Proof of Theorem 5.5 | 142 |
| | G.6 Proof of Theorem 5.6 | 143 |
| | G.7 Proof of Theorem 5.7 | 144 |
| | G.8 Proof of Theorem 5.8 | 145 |
| | G.9 Proof of Theorem 5.9 | 147 |
| | G.10 Proof of Theorem 5.10 | 148 |
| | G.11 Proof of Theorem 5.11 | 149 |
| | G.12 Proof of Theorem 5.12 | 150 |
| H | Loss Decomposition Results by Method | 151 |
| I | Loss Decomposition Results by Variable | 158 |
| J | Sanity Check for Experimental Results | 164 |
| | Bibliography | 173 |

List of Tables

| | | |
|------|--|-----|
| 4.1 | Datasets used in the experiments. | 25 |
| 4.2 | Ensemble loss comparison summary. | 37 |
| 4.3 | Averages from bias-variance decomposition (absolute values). | 48 |
| 4.4 | Averages from bias-variance decomposition (ratios relative to the base classifier). | 48 |
| 5.1 | Example where there is no diversity ($\overline{D} = 0$). | 62 |
| 5.2 | Example with high diversity and perfect classification. | 62 |
| 5.3 | Example with high diversity but no performance gain. | 62 |
| 5.4 | Example with high diversity and performance loss. | 62 |
| 6.1 | Loss comparison of <i>Bagging</i> (1; 30) with Probabilistic vs. Majority Vote. | 65 |
| 6.2 | Loss decomposition components for <i>Bagging</i> (1; 30) with Probabilistic vs. Majority Vote. | 66 |
| 6.3 | Comparison of sampling schemes. | 67 |
| B.1 | Performance of the base classifier. | 85 |
| F.1 | BVD results for <i>Bagging</i> (1; 30): absolute values. | 124 |
| F.2 | BVD results for <i>Bagging</i> (1; 30): relative ratios. | 125 |
| F.3 | BVD results for <i>Bagging</i> (0.5; 30): absolute values. | 126 |
| F.4 | BVD results for <i>Bagging</i> (0.5; 30): relative ratios. | 127 |
| F.5 | BVD results for <i>Bagging</i> (2; 30): absolute values. | 128 |
| F.6 | BVD results for <i>Bagging</i> (2; 30): relative ratios. | 129 |
| F.7 | BVD results for <i>Cragging</i> (2; 15): absolute values. | 130 |
| F.8 | BVD results for <i>Cragging</i> (2; 15): relative ratios. | 131 |
| F.9 | BVD results for <i>Cragging</i> (3; 10): absolute values. | 132 |
| F.10 | BVD results for <i>Cragging</i> (3; 10): relative ratios. | 133 |

| | | |
|------|---|-----|
| F.11 | BVD results for <i>Cragging</i> (30; 1): absolute values. | 134 |
| F.12 | BVD results for <i>Cragging</i> (30; 1): relative ratios. | 135 |
| F.13 | BVD results for the base classifier: absolute values. | 136 |
| | | |
| H.1 | Loss decomposition results for <i>Bagging</i> (1; 30) with Majority Vote. . . | 151 |
| H.2 | Loss decomposition results for <i>Bagging</i> (1; 30). | 152 |
| H.3 | Loss decomposition results for <i>Bagging</i> (0.5; 30). | 153 |
| H.4 | Loss decomposition results for <i>Bagging</i> (2; 30). | 154 |
| H.5 | Loss decomposition results for <i>Cragging</i> (2; 15). | 155 |
| H.6 | Loss decomposition results for <i>Cragging</i> (3; 10). | 156 |
| H.7 | Loss decomposition results for <i>Cragging</i> (30; 1). | 157 |
| | | |
| I.1 | Comparison of L | 158 |
| I.2 | Comparison of \bar{L} | 159 |
| I.3 | Comparison of \bar{D} | 160 |
| I.4 | Comparison of \bar{D}_T | 161 |
| I.5 | Comparison of \bar{D}_F | 162 |
| I.6 | Comparison of \bar{D}_P | 163 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Pseudo-code procedure LEARN. | 8 |
| 2.2 | Pseudo-code procedures SAMPLE, TRAIN, and TEST. | 9 |
| 2.3 | A taxonomy of ensemble learning methods. | 11 |
| 2.4 | The Bagging algorithm. | 14 |
| 2.5 | The Cragging algorithm. | 15 |
| 3.1 | Averaged error curves for $Cragging(n; 1)$ | 20 |
| 3.2 | Loss comparison for ensembles with 30 classifiers. | 21 |
| 4.1 | Error curves for $\mathcal{B}(0.5; n)$ and $\mathcal{B}(2; n)$ versus $\mathcal{B}(1; n)$ | 26 |
| 4.2 | Error curves for $C(2; n)$ and $C(3; n)$ versus $\mathcal{B}(1; n)$ | 32 |
| 4.3 | κ -Error Diagram summary – mean pairwise errors versus ensemble losses. | 40 |
| 4.4 | κ -Error Diagram summary – kappa values versus ensemble losses. | 41 |
| 4.5 | κ -Error Diagram summary. | 41 |
| 4.6 | Average margins versus average loss on test data. | 44 |
| 5.1 | Intuitive relations between L , \bar{L} , \bar{D} and y , \hat{y} and \hat{y}_c | 55 |
| 5.2 | Case study of relations between L , \bar{L} , and \bar{D} | 61 |
| 6.1 | Influence of decomposition variables on ensemble loss for $Bagging(0.5; 30)$, $Bagging(1; 30)$, and $Bagging(2; 30)$ | 68 |
| 6.2 | Influence of decomposition variables on ensemble loss for $Cragging(2; 15)$, $Cragging(3; 10)$, and $Cragging(30; 1)$ | 68 |
| 6.3 | Measured values of L , \bar{L} , and \bar{D} | 69 |
| 6.4 | Measured values of \bar{D}_T , \bar{D}_F , and \bar{D}_P | 70 |
| 6.5 | Values of $\bar{L}/(1 - \bar{D}_P - \bar{D}_F)$ and $\bar{D}/(1 - \bar{D}_P - \bar{D}_F)$ for $Bagging(0.5; 30)$, $Bagging(1; 30)$, and $Bagging(2; 30)$ | 71 |

| | | |
|-----|--|-----|
| 6.6 | Values of $\bar{L}/(1 - \bar{D}_P - \bar{D}_F)$ and $\bar{D}/(1 - \bar{D}_P - \bar{D}_F)$ for <i>Cragging</i> (2; 15), <i>Cragging</i> (3; 10), and <i>Cragging</i> (30; 1). | 72 |
| 6.7 | Consistency of changes in L , $\bar{L}/(1 - \bar{D}_P - \bar{D}_F)$, and $\bar{D}/(1 - \bar{D}_P - \bar{D}_F)$ when switching sampling schemes. | 74 |
| 6.8 | Relative ensemble losses and \bar{D}_P vs. number of classes. | 75 |
| C.1 | Error curves for <i>Cragging</i> (n ; 1). | 86 |
| D.1 | κ -Error Diagrams for $\mathcal{B}(0.5; 30)$, $\mathcal{B}(1; 30)$, and $\mathcal{B}(2; 30)$ | 88 |
| D.2 | κ -Error Diagrams for $C(2; 15)$, $C(3; 10)$, and $C(30; 1)$ | 95 |
| E.1 | Cumulative margin distributions for $\mathcal{B}(0.5; 30)$ | 103 |
| E.2 | Cumulative margin distributions for $\mathcal{B}(1; 30)$ | 106 |
| E.3 | Cumulative margin distributions for $\mathcal{B}(2; 30)$ | 109 |
| E.4 | Cumulative margin distributions for $C(2; 15)$ | 112 |
| E.5 | Cumulative margin distributions for $C(3; 10)$ | 115 |
| E.6 | Cumulative margin distributions for $C(30; 1)$ | 118 |
| E.7 | Cumulative margin distributions for the base classifier. | 121 |
| J.1 | Loss comparison for <i>Bagging</i> (0.5; 30), <i>Bagging</i> (1; 30), <i>Bagging</i> (2; 30), and the base classifier. | 165 |
| J.2 | Loss comparison for <i>Cragging</i> (2; 15), <i>Cragging</i> (3; 10), <i>Cragging</i> (30; 1), and the base classifier. | 168 |

Abbreviations

| | |
|----------|---|
| BVD | Bias-Variance Decomposition |
| Bagging | Bootstrap Aggregating |
| Cragging | Cross-Validation Aggregating |
| ECOC | Error-Correcting Output Coding |
| i.i.d. | independently and identically distributed |
| iff | if and only if |
| MCMC | Markov Chain Monte Carlo |
| vs. | versus |
| w.r.t. | with respect to |

Notation

| | |
|--|---|
| $a = b$ | Equality of a and b |
| $a := b$ | Definition of a as b ; or assignment of a to value of b |
| \forall | For all (universal quantifier) |
| $f : A \rightarrow B$ | Function f from A to B |
| $I()$ | Indicator function |
| $\{a_1, \dots, a_n\}$ | Finite set or multi-set consisting of n elements a_1, \dots, a_n |
| $\{a \mathcal{P}(a)\}$ | Set containing all elements a for which $\mathcal{P}(a)$ is true |
| \emptyset | Empty set |
| $ A $ | Cardinality (number of elements in set or tuple A) |
| $ a $ | Absolute value of numeric variable a |
| $A \times B$ | Cartesian product of two sets A and B |
| A^n | $A \times A \times \dots \times A$ (n times) iff A is a set, $A * A * \dots * A$ (n times) iff A is a number |
| \mathcal{R} | The set of all real numbers |
| \mathcal{N} | The set of all natural numbers, including 0 |
| \mathcal{N}^+ | The set of all positive natural numbers |
| ∞ | Infinity |
| $\mathbf{a} = \langle a_1, \dots, a_n \rangle$ | Ordered tuple \mathbf{a} consisting of n elements a_1, \dots, a_n |
| $p(a)$ | Probability that a is true iff a is a predicate |
| $P(A)$ | Probability that $A = a$ iff A is a random variable |
| $P(A, B)$ | Probability distribution of random variable A |
| $P(A B)$ | Joint probability distribution of random variables A and B |
| $P(A b)$ | Probability distribution of A conditioned B |
| $\{P(A)\}$ | Probability distribution of A , given that $B = b$ |
| $\{P(\mathcal{R})\}$ | Set containing all probability distributions $P(A)$ |
| $E_{a \in A} [V_a]$ | Set containing all probability distributions over \mathcal{R} |
| $E_{P(c)} [V_c]$ | Expectation of random variable V taken over set A |
| \mathbf{X} | Expectation of random variable V according to $P(c)$ |
| Y | Input space |
| \hat{Y} | Outcome space |
| \hat{Y}_c | Ensemble prediction space |
| $\mathbf{s} = \langle \mathbf{x}, y \rangle$ | Member prediction space |
| $\mathbf{S} = \langle \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m \rangle$ | A given input/outcome - pair (instance) |
| $P(\mathbf{X}, Y)$ | Multi-set (sample) of instances |
| | Probability distribution over instances |

| | |
|--|---|
| $p(\mathbf{x}, y)$ | Probability of encountering instance $\langle \mathbf{x}, y \rangle$ |
| C | Classifier or ensemble $C : \mathbf{X} \rightarrow \hat{Y}$ |
| J | Classifier inducer (learner) |
| l | Loss function $l : \hat{Y} \times Y \rightarrow \mathcal{R}$ |
| $l(\hat{y}, y)$ | Loss of prediction \hat{y} relative to outcome y |
| $l_2(\hat{y}, y)$ | Squared loss of prediction \hat{y} relative to outcome y |
| $l_{01}(\hat{y}, y)$ | 0-1 loss of prediction \hat{y} relative to outcome y |
| $l_{ }(\hat{y}, y)$ | absolute loss of prediction \hat{y} relative to outcome y |
| k | number of classes in discrete outcome space |
| m | sample size for sample of instances |
| n | number of classifiers in ensemble |
| c, C_i | Member classifiers |
| V | Voting function |
| $\mathbf{c} = \langle c_1, \dots, c_n \rangle$ | Tuple of member classifiers |
| $\mathbf{w} = \langle w_1, \dots, w_n \rangle$ | Tuple of voting weights |
| $\hat{y}_C(\mathbf{x})$ | Classifier prediction |
| $\hat{y}(\mathbf{x})$ | Ensemble prediction |
| $\hat{\mathbf{y}}(\mathbf{x})$ | Tuple of ensemble members' predictions |
| $\hat{y}_c(\mathbf{x})$ | Ensemble member prediction |
| $\hat{P}(Y x)$ | Belief distribution of probabilistic classifier for input x |
| $R(\hat{y} x)$ | Conditional risk of predicting \hat{y} for input x |
| $L(\mathbf{x}, y)$ | Loss of ensemble for instance $\langle \mathbf{x}, y \rangle$ |
| $\bar{L}(\mathbf{x}, y)$ | Mean member loss of ensemble for instance $\langle \mathbf{x}, y \rangle$ |
| $\bar{D}(\mathbf{x})$ | Diversity of ensemble members for input x |
| L | Expected loss of ensemble for domain |
| \bar{L} | Expected mean member loss of ensemble for domain |
| \bar{D} | Expected diversity of ensemble for domain |
| $\mathcal{B}(s; n)$ | <i>Bagging</i> ($s; n$) (using n runs and relative sample size s) |
| $C(f; n)$ | <i>Cragging</i> ($f; n$) (using n runs and f folds) |