

Dataset Selection for Aggregate Model Implementation in Predictive Data Mining

by

Patricia Elizabeth Nalwoga Lutu

Thesis submitted in partial fulfilment of the requirements for the
degree of

Philosophiae Doctor

in the Faculty of Engineering Built Environment and Information
Technology

The University of Pretoria

Pretoria

September 2010

Title: Dataset Selection for Aggregate
Model Implementation in Predictive
Data Mining

Author: Patricia Elizabeth Nalwoga Lutu

Abstract

Data mining has become a commonly used method for the analysis of organisational data, for purposes of summarizing data in useful ways and identifying non-trivial patterns and relationships in the data. Given the large volumes of data that are collected by business, government, non-government and scientific research organizations, a major challenge for data mining researchers and practitioners is how to select relevant data for analysis in sufficient quantities, in order to meet the objectives of a data mining task. This thesis addresses the problem of dataset selection for predictive data mining. Dataset selection was studied in the context of aggregate modeling for classification.

The central argument of this thesis is that, for predictive data mining, it is possible to systematically select many dataset samples and employ different approaches (different from current practice) to feature selection, training dataset selection, and model construction. When a large amount of information in a large dataset is utilised in the modeling process, the resulting models will have a high level of predictive performance and should be more reliable. Aggregate classification models, also known as ensemble classifiers, have been shown to provide a high level of predictive accuracy on small datasets. Such models are known to achieve a reduction in the bias and variance components of the prediction error of a model. The research for this thesis was aimed at the design of aggregate models and the selection of training datasets from large amounts of available data. The objectives for the model design and dataset selection were to reduce the bias and variance components of the prediction error for the aggregate models.

Design science research was adopted as the paradigm for the research. Large datasets obtained from the UCI KDD Archive were used in the experiments. Two classification algorithms: See5 for classification tree modeling and K-Nearest

Neighbour, were used in the experiments. The two methods of aggregate modeling that were studied are One-Vs-All (OVA) and positive-Vs-negative (pVn) modeling. While OVA is an existing method that has been used for small datasets, pVn is a new method of aggregate modeling, proposed in this thesis. Methods for feature selection from large datasets, and methods for training dataset selection from large datasets, for OVA and pVn aggregate modeling, were studied.

The experiments of feature selection revealed that the use of many samples, robust measures of correlation, and validation procedures result in the reliable selection of relevant features for classification. A new algorithm for feature subset search, based on the decision rule-based approach to heuristic search, was designed and the performance of this algorithm was compared to two existing algorithms for feature subset search. The experimental results revealed that the new algorithm makes better decisions for feature subset search. The information provided by a confusion matrix was used as a basis for the design of OVA and pVn base models which are combined into one aggregate model. A new construct called a *confusion graph* was used in conjunction with new algorithms for the design of pVn base models. A new algorithm for combining base model predictions and resolving conflicting predictions was designed and implemented. Experiments to study the performance of the OVA and pVn aggregate models revealed the aggregate models provide a high level of predictive accuracy compared to single models. Finally, theoretical models to depict the relationships between the factors that influence feature selection and training dataset selection for aggregate models are proposed, based on the experimental results.

Key words:

data mining, predictive modeling, classification, model aggregation, ensemble classifiers, OVA classification, pVn classification, dataset selection, feature selection, variable selection, bias reduction, variance reduction, large datasets, dataset sampling, dataset partitioning.

Thesis supervisor: Prof. A.P. Engelbrecht
Department: Department of Computer Science
Degree: Philosophiae Doctor
(Doctor of Philosophy)

Dedication

This thesis is dedicated in loving memory to my parents

Omwami Wilson Sebowa Lutu

and

Omumbejja Kasalina Zalwango Lutu.

Acknowledgements

I wish to express my sincere gratitude to my research supervisor Prof. Andries Engelbrecht. Thank you for all the time you have dedicated to advising me on my research activities, and especially for reading my thesis with what I call the magnifying glass! I have been privileged to benefit from your extensive research and authoring experience.

I also wish to express my sincere gratitude to the external examiners for taking the time to examine this thesis. Thank you for the encouraging feedback on the thesis contents, and the constructive advice you have given for the final thesis revisions.

I wish to thank the following people: Prof. Carina de Villiers, head, Department of Informatics. Thank you for all the support you have given me over the last five years, especially the research leave. My colleagues in the Statistics department: Dr. Raphael Kasonga, Mev. Judy Coetsee, and Mev. Dorothea Corbett. Thank you for the numerous advice you have given me on statistical inference. My colleagues in the library: Mnr. Danie Malan, Ma. Tebogo Mogakane, and Mev. Gerda Ehlers. Thank you for all the assistance you have given me in acquiring research articles and books.

I also wish to thank all the people and angels that have given me spiritual guidance over the years. Prof. Ojelanki Ngwenyama: Thank you for the spiritual teachings and for introducing me to Toulmin's work.

Last but not least, I wish to thank my son Subi for patiently supporting and encouraging me with my research. You often asked me: 'Mummy, when are you going to finish that thesis?' Well Subi, after five and half years, I have completed the thesis.

Contents

1	Introduction	1
1.1	Motivation for the research.....	1
1.2	Current debates and practices in data mining from large datasets	3
1.3	Scope of the research	5
1.4	The claims of the thesis.....	6
1.5	Research paradigm	10
1.6	Research contributions.....	11
1.6.1	Methods and instantiations	12
1.6.2	Constructs, models and better theories	13
1.7	Overview of the thesis	13
2	Dataset Selection and Modeling from Large Datasets	16
2.1	The need for dataset selection.....	16
2.1.1	Customer Relationship Management - CRM.....	17
2.1.2	Web usage mining and electronic commerce.....	18
2.1.3	Forensic data mining.....	18
2.1.4	Scientific applications of data mining.....	19
2.2	Classification modeling from very large datasets.....	20
2.2.1	Terminology for classification modeling.....	21
2.2.2	The classification modeling problem.....	23
2.2.3	Single model construction.....	24
2.2.4	Aggregate model construction	25
2.2.5	Serial and parallel model aggregation	28
2.2.6	Model testing.....	30
2.3	The dataset selection problem	31
2.4	Theoretical methods for single sample selection	32
2.4.1	Probably Approximately Correct (PAC) learning	33
2.4.2	The Hoeffding-Chernoff bounds	34
2.5	Empirical methods for single sample selection	35
2.5.1	The Dynamic Sampling method.....	35
2.5.2	The progressive sampling method.....	36
2.5.3	Static sample size estimation	37
2.5.4	Density-biased sampling.....	37
2.5.5	One-sided sampling	37
2.6	Methods for selecting multiple training datasets	38
2.6.1	Bootstrap sampling and boosting of small datasets	39
2.6.2	Partitioning of large datasets	40
2.6.3	Combining dataset sampling and partitioning.....	41



2.7 Conceptual views of classification modeling	41
2.7.1 Discriminative classification	42
2.7.2 Probabilistic classification	42
2.7.3 Definition of decision boundaries and class confusion regions	43
2.7.4 Selection of training data to support the objectives of classification	44
2.8 Sources of classification error	45
2.8.1 Bias, variance and intrinsic errors in classification	46
2.8.2 Factors that influence the components of prediction error	47
2.8.3 Selection of training data to reduce classification error	49
2.9 The limitations of current methods of dataset selection	49
2.10 Proposed approach to selection of training data from very large datasets	50
2.10.1 Variance reduction methods	51
2.10.2 Bias reduction methods	51
2.11 Conclusions	52
3 The Feature Selection Problem	53
3.1 The need for feature selection	53
3.1.1 Feature relevance and redundancy	54
3.1.2 The curse of dimensionality	55
3.2 Implicit feature selection	55
3.3 Explicit feature selection	56
3.3.1 Categories of feature selection methods	56
3.3.2 Feature selection using wrapper methods	57
3.3.3 Feature selection based on pure ranking	57
3.3.4 Feature selection based on heuristic search	58
3.3.5 Feature selection using relevance and redundancy analysis	59
3.3.6 Feature selection for large datasets	59
3.4 Merit measures for heuristic search of feature subsets	60
3.5 Measuring correlations	63
3.5.1 Problems with Pearson's correlation coefficient	63
3.5.2 Robust measures of correlation	64
3.6 Validation methods for feature selection	65
3.6.1 The need for validation of correlation coefficients	66
3.6.2 Practical significance of correlation coefficients	66
3.6.3 Validation based on hypothesis testing for correlation coefficients	67
3.6.4 Validation based on fake variables	68
3.7 Conclusions	69
4 Research Methods	71
4.1 Research questions and objectives	71

4.2	The central argument for the thesis	72
4.3	The research paradigm and methodology	73
4.3.1	The design science research paradigm.....	73
4.3.2	The outputs of design science research.....	75
4.3.3	Artifact evaluation and theory building.....	75
4.3.4	Justification for adopting the design science research paradigm.....	77
4.3.5	Theories for data mining	78
4.4	The datasets used in the experiments	78
4.4.1	Choice of datasets and past usage	79
4.4.2	Dataset pre-processing to balance class distributions	82
4.4.3	Dataset pre-processing to normalise feature values	85
4.5	Sampling methods.....	86
4.5.1	Sequential random sampling	87
4.5.2	Obtaining random samples from datasets	87
4.6	The data mining algorithms used in the experiments	87
4.6.1	Classification trees.....	88
4.6.2	K-Nearest Neighbour classification.....	89
4.7	Measures of model performance	90
4.7.1	Measures of predictive performance	90
4.7.2	Statistical test to compare model performance	92
4.7.3	Analysis of performance using ROC curves and lift charts	94
4.8	Software used for the experiments	97
4.9	Chapter summary.....	98
5	Feature Selection for Large Datasets.....	100
5.1	The feature selection problem revisited	101
5.2	Alternative approaches to feature selection for large datasets	102
5.3	Empirical study of feature ranking methods for large datasets	104
5.3.1	Experimental procedure for the study of feature ranking.....	104
5.3.2	Comparison of Pearson's and Kendall's correlation measures.....	105
5.3.3	Feature ranking based on a single sample.....	108
5.3.4	Feature ranking based on many samples	109
5.4	Empirical study of feature subset search	111
5.4.1	Implementation of feature relevance and redundancy definitions	112
5.4.2	A reliable search procedure for feature subset search.....	116
5.5	Predictive performance of features selected with different methods	123
5.5.1	Experimental procedure for classifier creation and testing.....	123
5.5.2	Classification results for forest cover type	125
5.5.3	Classification results for KDD Cup 1999.....	130
5.5.4	Classification results for the small datasets.....	133



5.6 Discussion	136
5.6.1 Correlation measures and feature ranking	136
5.6.2 Feature subset selection.....	138
5.6.3 Problems associated with the global measurement of correlations	139
5.7 Conclusions.....	139
6 Methods for Dataset Selection and Base Model Aggregation	141
6.1 Problem decomposition for OVA and pVn modeling.....	142
6.1.1 Problem decomposition for OVA modeling.....	143
6.1.2 Problem decomposition for pVn modeling	143
6.2 Methods for improving predictive performance.....	144
6.2.1 Reduction of bias and variance errors for small datasets.....	144
6.2.2 Reduction of bias and variance errors for large datasets	145
6.2.3 High competence and syntactic diversity of base models	146
6.3 Design and selection of training and test datasets	147
6.3.1 Strategy for dataset selection and model creation	148
6.3.2 Motivation for the sampling methods.....	148
6.3.3 Partitioning and sampling for dataset selection	150
6.3.4 Sampling from dataset partitions	151
6.4 Methods for creating and testing OVA and pVn models	152
6.4.1 Design and implementation of OVA and pVn base models	152
6.4.2 Implementation of OVA and pVn aggregate models	154
6.4.3 Algorithms for model aggregation.....	155
6.4.4 Experimental procedure for testing aggregate models.....	158
6.4.5 Measurement of performance gains for OVA and pVn aggregate models.....	159
6.5 Chapter summary.....	161
7 Evaluation of Dataset Selection for One-Versus-All Aggregate	
Modeling	162
7.1 OVA modeling	162
7.1.1 Motivation for OVA modeling	163
7.1.2 Sample composition for OVA base model training datasets	163
7.1.3 Experiment design for the study of OVA modeling.....	164
7.2 Experiments to study OVA models for 5NN classification	164
7.2.1 Predictive performance of un-boosted 5NN OVA models	165
7.2.2 Design of boosted 5NN OVA base models	168
7.2.3 Predictive performance of boosted 5NN OVA models	171
7.3 Experiments to study OVA models for See5 classification	175
7.3.1 Predictive performance of un-boosted See5 OVA models.....	175
7.3.2 Design of See5 boosted OVA base models	178

7.3.3 Predictive performance of boosted See5 OVA models	179
7.4 Discussion	183
7.5 Conclusions	184
8 Evaluation of Dataset Selection for Positive-Versus-Negative	
Aggregate Modeling	186
8.1 pVn modeling	187
8.1.1 Motivation for pVn modeling	187
8.1.2 Design of pVn base models.....	187
8.1.3 Experiment design for the study of pVn modeling	188
8.2 Experiments to study pVn models for 5NN classification.....	189
8.2.1 Design of training datasets for 5NN pVn base models.....	189
8.2.2 Predictive performance of the 5NN pVn base models.....	193
8.2.3 Predictive performance of the 5NN pVn aggregate models	194
8.3 Experiments to study pVn models for See5 classification	196
8.3.1 Design of training datasets for pVn base models	197
8.3.2 Predictive performance of the See5 pVn base models	200
8.4 Comparison of performance variability for single and aggregate models.....	203
8.5 Discussion	205
8.5.1 Dataset selection for pVn modeling.....	205
8.5.2 Comparison of OVA and pVn modeling.....	206
8.5.3 Classification problems where proposed boosting methods are not appropriate	207
8.6 Conclusions	209
9 ROC Analysis for Single and Aggregate Models	211
9.1 ROC analysis for 2-class predictive models.....	212
9.2 ROC analysis for multi-class predictive models.....	212
9.3 ROC analysis for 5NN models	214
9.4 ROC analysis for See5 models.....	216
9.5 Conclusions	218
10 Recommendations for Dataset Selection	220
10.1 Reduction of prediction error.....	220
10.2 Recommendations for feature selection	221
10.2.1 Summary of the feature selection experimental results.....	221
10.2.2 Guidelines for feature selection	223
10.3 Recommendations for training dataset selection for aggregate modeling.....	225
10.3.1 Summary of the training dataset selection experimental results	225
10.3.2 Theoretical model for training dataset selection	226
10.3.3 Parallel versus serial aggregation of base models	228

10.3.4 Guidelines for OVA and pVn model design, training dataset selection and testing	229
10.5 Chapter summary	231
11 Discussion of Research Contributions	232
11.1 Outputs of design science research	232
11.2 Evaluation of design science research	233
11.2.1 Criteria for design science research evaluation	233
11.2.2 Constructs, models and better theories	234
11.2.3 Methods and instantiations	235
11.2.4 Rigorous design evaluation	235
11.2.5 Rigor and design as a search process	237
11.2.6 Research contributions for design science research	238
11.3 Limitations of the proposed dataset selection methods	239
11.4 Chapter Summary	240
12 Conclusions	241
12.1 Summary of the thesis	241
12.2 Conclusions and reflection	242
12.3 Future work	243
References	245
Appendices	261
A Definition of symbols	262
B Definitions of statistical measures	265
C Descriptive statistics for the datasets	270
D Correlation measurements for feature selection	277
E Algorithm for breadth first generation of a search space	283
F Predictive performance for single OVA and pVn models	285
G ROC analysis details	300
H Using statistical and database software to implement dataset selection methods	308
I Publications and conference presentations	309

List of Figures

2.1	A typical learning curve	25
2.2	Confusion region for two classes	44
2.3	Components of prediction error and factors that influence prediction error	48
4.1	A general model for generating knowledge in design science research.....	74
4.2	Relationship between understanding, generalisation and scientific theories.....	76
4.3	Steps of the scientific method.....	77
4.5	ROC space and AUC	95
5.1	Merit values for the forest cover type dataset without pre-selection	115
5.2	Merit values for the KDD Cup 1999 dataset without feature pre-selection	115
5.3	Decision rule-based algorithm based on definitions of relevance and redundancy	119
5.4	The algorithm GetBestInCat(CT) to select the best features in one category	120
6.1	Steps for dataset partitioning, model creation and testing	148
6.2	Partitioning and sampling process for base model training dataset selection	150
6.3	Algorithm for combining See5 base model predictions	156
6.4	Algorithm for combining 5NN base model predictions	157
6.5	Experimental method for aggregate model implementation for one test set.....	159
8.1	Confusion graph for the 5NN single 7-class model for Forest cover type for training set size of 12000 instances	190
8.2	Confusion graph for the 5NN single 5-class model for KDD Cup 1999 for training set size of 4000 instances	190
8.3	Algorithm for class selection for the pVn base models	191
8.4	Confusion graph for the See5 single 7-class model for forest cover type for training set size of 12000 instances	198
8.5	Confusion graph for the See5 single 5-class model for KDD Cup 1999 for training set size of 4000 instances	198
8.6	Modified algorithm for class selection for the pVn base models	199
8.7	Simplified confusion graph for the See5 single 5-class model for KDD Cup 1999	199
10.1	Theoretical predictive model for feature selection using filtering methods	223
10.2	Recommended procedure for feature selection from large datasets	224
10.3	Theoretical predictive model for aggregate model performance based on existing literature	227
10.4	Extensions to the theoretical predictive model for aggregate model performance based on studies for this thesis	228
10.5	Steps for the creation of a confusion matrix and confusion graph	229
10.6	Steps for the design, creation and testing of un-boosted OVA aggregate models	230
10.7	Steps for the design, creation and testing of boosted OVA aggregate models	230
10.8	Steps for the design, creation and testing of pVn aggregate models	231



C.1	Class frequencies for the forest cover type class variable (covertype)	270
C.2	Class frequencies for the KDD Cup 1999 training dataset derived class variable (class)	271
C.3	Class frequencies for the abalone3C class variable (age)	273
C.4	Class frequencies for the wine quality (white) class variable (quality).....	274
E.1	Breadth-first search algorithm	283
E.2	BreadthFirstGenerate algorithm.....	284
G.1	Areas of the ROC plane used to compute the AUC	300

List of Tables

4.1	Outputs of design science research	75
4.2	Examples of datasets used in data mining and machine learning studies.....	79
4.3	The datasets used for the experiments	81
4.4	Class counts for the forest cover type dataset	81
4.5	Class counts for the KDD Cup 1999 training(10% version) and test sets.....	82
4.6	Reduction of the over-representation of (service, attack type) values in the KDD Cup 1999 training and test datasets	83
4.7	Class counts for the final version of the KDD Cup 1999 training dataset	84
4.8	Class counts for the final version of the KDD Cup 1999 test dataset	85
4.9	Range of values for features in the KDD Cup 1999 dataset	86
4.9	Theoretical confusion matrix for a 2-class model.....	91
4.10	Measures of performance derived from a confusion matrix	91
4.11	Interpretation of p values for statistical tests	94
4.12	Software used for the experiments.....	98
5.1	Characteristics of the probes for the datasets.....	105
5.2	Comparison of mean values for Kendall's tau and Pearson's r	106
5.3	Comparison of the number of selected features for Kendall's tau and Pearson's r	108
5.4	Number of selected features based on single samples for forest cover type	109
5.5	Kendall's correlations for four features for KDD Cup 1999	109
5.6	Number of selected features based on 10 samples.....	110
5.7	Interpretation of levels of feature correlations for heuristic search.....	113
5.8	Trace of the CFS search procedure for the forest cover type and KDD Cup 1999.....	114
5.9	Proposed definition of feature relevance and redundancy based on user specified levels	118
5.10	Decision rules for choosing between two features of the same category	120
5.11	Output of the decision rule-based search algorithm without feature pre-selection for KDD Cup 1999.....	121
5.12	Output of the decision rule-based search algorithm without feature pre-selection for forest cover type	122
5.13	Features selected by the decision rule-based algorithm for sample sizes of 1000.....	123
5.14	Predictive accuracy for forest cover type based on two class distributions	125
5.15	Statistical tests to compare the accuracy of forest cover type classifiers for different feature subsets for parent dataset class distribution	126
5.16	Statistical tests to compare the accuracy of forest cover type classifiers for different feature subsets for equal class distribution	128
5.17	Statistical tests to compare TPRATE performance of forest cover type classifiers for different feature subsets for training sample size 12000.....	129
5.18	Predictive performance of KDD Cup 1999	131

5.19	Statistical tests to compare the performance of KDD Cup 1999 classifiers for different feature subsets	132
5.20	Predictive accuracy for the small datasets based on the parent dataset class distribution	134
5.21	Statistical tests to compare the predictive performance of small dataset classifiers ..	135
6.1	Interpretation of Ali and Pazzani (1996) measures	160
7.1	Predictive performance of 5NN OVA un-boosted base models	165
7.2	Predictive performance of 5NN single and un-boosted OVA aggregate models	166
7.3	Statistical tests to compare the performance of 5NN single and un-boosted OVA aggregate models for forest cover type	167
7.4	Statistical tests to compare the performance of 5NN single and un-boosted OVA aggregate models for KDD Cup 1999	168
7.5	Confusion matrix for the 5NN single model for the forest cover type dataset.....	169
7.6	5NN training sample composition to reduce class confusion for forest cover type.....	170
7.7	Confusion matrix for the 5NN single model for the KDD Cup 1999 dataset	170
7.8	Training sample composition to reduce class confusion for 5NN models for KDD Cup 1999.....	171
7.9	Predictive performance of 5NN OVA boosted base models	172
7.10	Predictive performance of 5NN single, un-boosted and boosted OVA aggregate models	172
7.11	Statistical tests to compare the 5NN single, un-boosted and boosted OVA aggregate models for forest cover type	173
7.12	Statistical tests to compare the 5NN single, un-boosted and boosted OVA aggregate models for KDD Cup 1999.....	174
7.13	Predictive performance of See5 OVA un-boosted base models.....	175
7.14	Predictive performance of See5 single and un-boosted OVA aggregate models.....	176
7.15	Statistical tests to compare the performance of See5 single and un-boosted OVA aggregate models for forest cover type.....	177
7.16	Statistical tests to compare the performance of See5 single and un-boosted OVA aggregate models for KDD Cup 1999	177
7.17	Confusion matrix for See5 classification tree single 7-class model for forest cover type	178
7.18	Confusion matrix for See5 classification tree single 5-class model for KDD Cup 1999	178
7.19	See 5 Training sample composition to reduce class confusion for KDD Cup 1999....	179
7.20	Predictive performance of See5 OVA boosted base models	179
7.21	Predictive performance of See5 single, un-boosted and boosted OVA aggregate models	180
7.22	Statistical tests to compare the See5 single, un-boosted and boosted OVA aggregate models for forest cover type	181

7.23	Statistical tests to compare the See5 single and boosted OVA aggregate models for KDD Cup 1999.....	183
7.24	Summary of the conclusions from the OVA modeling experiments.....	184
7.25	Sample of the output for the See5 combination algorithm	184
8.1	Trace of the class selection algorithm for the 5NN forest cover type graph	192
8.2	5NN training set composition for the pVn base models for forest cover type and KDD Cup 1999	193
8.3	Predictive performance of 5NN pVn base models	193
8.4	Mean Predictive performance of the 5NN single, OVA and pVn aggregate models for forest cover type	194
8.5	Statistical tests to compare the performance for 5NN single and pVn aggregate models for forest cover type.....	195
8.6	Mean Predictive performance of single, OVA and pVn aggregate 5NN models for KDD Cup 1999	196
8.7	Statistical tests to compare the 5NN single and pVn aggregate models for KDD Cup 1999	196
8.8	Training set composition for the See5 pVn base models.....	200
8.9	Predictive performance of See5 pVn base models	200
8.10	Predictive performance of the See5 single, OVA and pVn models for forest cover type	201
8.11	Statistical tests to compare the performance for See5 classification tree single and pVn aggregate models for forest cover type.....	202
8.12	Predictive performance of See5 single, OVA and pVn aggregate models for KDD Cup 1999.....	203
8.13	Statistical tests to compare See5 single and pVn aggregate models for KDD Cup 1999	203
8.14	F- tests for comparison of performance variability for single and aggregate models..	204
8.15	Summary of performance improvements for OVA and pVn models	206
8.16	See5 single 3-class model confusion matrix for abalone3C	208
8.17	See5 single 3-class model confusion matrix for waveform	209
9.1	Computations for the estimation of the VUS	214
9.2	ROC analysis results for the 5NN single and aggregate models	215
9.3	ROC analysis results for the See5 single and aggregate models.....	217
11.1	Criteria for the evaluation of design science research	234
11.2	Summary of new algorithms	238
	of appendices	261
A.1	Symbols used in the thesis	262
C.1	Descriptive statistics for the quantitative variables in the forest cover type dataset....	270
C.2	Descriptive statistics for the qualitative variables for the forest cover type dataset	271

C.3	Descriptive statistics for the quantitative variables for the KDD Cup 1999 training dataset	272
C.4	Descriptive statistics for the qualitative variables for the KDD Cup 1999 training dataset	273
C.5	Descriptive statistics for the quantitative variables of abalone3C.....	274
C.6	Descriptive statistics for the Wine quality (white) dataset variables	275
C.8	Descriptive statistics for the mushroom dataset variables.....	276
D.1	Feature selection for Forest cover type	277
D.2	Feature selection for forest cover type using Kendall's tau and a Gaussian probe	278
D.3	Features selected by the decision rule-based search algorithm for different inputs	279
D.4	Feature selection for KDD Cup 1999	279
D.5	Feature selection for KDD Cup 1999 using Kendall's tau and the Gaussian probe....	280
D.6	KDD Cup 1999 feature selection by decision rule	281
D.7	Feature selection for Abalone using Pearson's r and Kendall's tau	281
D.8	Abalone3C feature-feature correlations	282
D9	Feature selection for mushroom using SU coefficients	282
F.1	Predictive performance of the 5NN single 7- class model for forest cover type	285
F.2	Predictive performance of the 5NN un-boosted OVA aggregate model for forest cover type	286
F.3	Predictive performance of the 5NN boosted OVA aggregate model for forest cover type	286
F.4	Predictive performance of the 5NN pVn aggregate model for forest cover type	287
F.5	Predictive performance of the See5 single 7-class model for forest cover type	287
F.6	Predictive performance of See5 un-boosted OVA aggregate model for forest cover type	288
F.7	Predictive performance of See5 boosted OVA aggregate model for forest cover type	288
F.8	Predictive performance of the See5 pVn aggregate model for forest cover type	289
F.9	Predictive performance of the 5NN single 5-class model for KDD Cup 1999.....	289
F.10	Predictive performance of the 5NN OVA un-boosted aggregate model for KDD Cup 1999.....	290
F.11	Predictive performance of the 5NN OVA boosted aggregate model for KDD Cup 1999	290
F.12	Predictive performance of the 5NN pVn aggregate model for KDD Cup 1999.....	291
F.13	Predictive performance of the See5 single model for KDD Cup 1999.....	291
F.14	Predictive performance of the See5 un-boosted OVA aggregate model for KDD Cup1999	292
F.15	Predictive performance of the See5 boosted OVA aggregate model for KDD Cup1999	292
F.16	Predictive performance of the See5 pVn aggregate model for KDD Cup 1999.....	293
F.17	Predictive performance of the 5NN single model for Wine quality.....	293

F.18	Predictive performance of the 5NN un-boosted OVA model for Wine quality	294
F.19	Predictive performance of the 5NN boosted OVA model for Wine quality.....	294
F.20	Predictive performance of the 5NN pVn model for Wine quality.....	295
F.21	Predictive performance of the See5 single model for Wine quality.....	295
F.22	Predictive performance of the See5 un-boosted model for Wine quality.....	296
F.23	Predictive performance of the See5 boosted model for Wine quality	296
F.24	Predictive performance of the See5 pVn model for Wine quality.....	297
F.25	Statistical tests for 5NN single and aggregate model comparison for wine quality	298
F.26	Statistical tests for See5 single and aggregate model comparison for wine quality ...	299
G.1	Method used for the computation of the AUC for probabilistic classifiers	301
G.2	One-vs-rest AUC for the 5NN forest cover type models	302
G.3	One-vs-rest AUC for the 5NN KDD Cup 1999 models	303
G.4	One-vs-rest AUC for the 5NN Wine quality models	304
G.5	One-vs-rest AUC for the See5 forest cover type models	305
G.6	One-vs-rest AUC for the See5 KDD Cup 1999 models.....	306
G.7	One-vs-rest AUC for the See5 Wine quality models	307
H.1	Suggestions for feature selection using statistical software	308
H.2	Suggestions for OVA and pVn modeling using statistical software.....	308



*The Infinite Intelligence
is beyond human understanding.*

*The Infinite Intelligence
created the universe:
all that we perceive,
and all that we do not perceive.*

*The Infinite Intelligence
exists in silence,
gives in silence,
and
loves in silence.*