References

AHA, D. W. & BANKERT, R. L. (1996) A comparative evaluation of sequential feature selection algorithms IN FUSHER, D. & LENZ, H. J. (Eds.) *Learning from Data: Artificial Intelligence and Statistics* Springer-Verlag.

ALI, K. M. & PAZZANI, J. (1996) Error reduction through learning multiple descriptions. *Machine Learning,* 24,173-202.

ASUNCION, A. & NEWMAN, D. J. (2007) UCI machine learning repository [http://0-www.ics.uci.edu./~mlearn/MLRepository.html]. Irvine, CA, University of California, Department of Information and Computer Science.

BAY, S. D., KIBLER, D., PAZZANI, M. J. & SMYTH, P. (2000) The UCI KDD archive of large data sets for data mining research and experimentation. *ACM SIGKDD,* 2 (2), 81-85.

BEKKERMAN, R., EL-YANIV, R., TISHBY, N. & WINTER, Y. (2003) Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research,* 3, 1183-1208.

BERRY, M.J.A & LINOFF, G.S. (2000) *Mastering Data Mining: The Art and Science of Customer Relationship Management.* John Wiley & Sons.

BI, J., BENNET, K. P., EMBRECHTS, M., BRENEMAN, C. M. & SONG, M. (2003) Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research,* 3, 1229-1243.

BISHOP, C. M. (1995) *Neural Network for Pattern Recognition.* Oxford:Clarendon Press.

BLACKARD, J. A. (1998) Comparison of Neural Network and Discriminant Analysis in Predicting Forest Cover Types, PhD Thesis. *Department of Forest Sciences.* Fort Collins, Colorado, Colorado State University.

BLAKE, C. L. & MERZ, C. J. (1998) UCI Repository of Machine Learning Databases. *Department of Computer Science.* Irvine, University of California.

BLUM, A. L. & LANGLEY, P. (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence,* 97(1-2), 245-271.

BOSE, R. (2002) Customer Relationship Management: Key components for IT success. *Industrial Management and Data Systems,* 102 (2). 89-97.

BOSER, B.E., GUYON, I.M. & VAPNIK, V.N . (1992) A training algorithm for optimal margin classifiers. In D. HAUSSLER, editor, *5th Annual ACM Workshop on COLT*, 144-152.

BREIMAN, L. (1996) Bagging predictors. *Machine Learning,* 24, 123-140.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984) *Classification and Regression Trees.* Pacific Grove, CA:Wadsworth & Brooks.

CATLETT, G. (1991) Megainduction: a test flight. *Proceeding of Eighth Workshop on Machine Learning.* San Mateo, California, Morgan Kaufmann.

CHAN, P. & STOLFO, S. (1998) Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.* AAAI

CHAWLA, N., MOORE, T. E., BOWYER, K. W., HALL, L. O., SPRINGER, C. & KEGELMEYER, P. (2001) Bagging is a small-data-set phenomenon. *In International Conference on Computer Vision and Pattern Recognition (CVPR), 2001.*

CLARK, D., SCHRETER, Z. & ADAMS, A. (1996) A quantitative comparison of distal and backpropagation. *Proceedings of the Seventh Australian Conference on Neural Networks (ACNN'96).* Canberra Australia.

COETSEE, J. (2007) *Private Communication.* Department of Statistics, University of Pretoria.

COHEN, J. (1988) *Statistical Power Analysis for the Behavioural Sciences, 2nd Edition.* Hillsdale: New Jersey Lawrence Erlbaum Associates.

COHEN, P.R. (1995) *Empirical Methods in Artificial Intelligence,* MIT Press: Cambridge, Massachusetts.

CORTEZ, P., CERDEIRA, F., ALMEIDA, F., MATOS, T. & REIS, J. (2009) Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems,* 42, 547-553.

COVER, T. M. & HART, P. E. (1967) Nearest Neighbor Pattern Classification. *IEEE Transaction on Information Theory,* IT-13 (1), 21-27.

DIETTERICH, T. (1995) Overfitting and Undercomputing in Machine Learning. *ACM Computing Surveys,* 27 (3), 326-327.

DIETTERICH, T. & BAKIRI, G. (1995) Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research,* 2, 263-286.

DIETTERICH, T. & KONG, E. (1995) Machine learning bias, statistical bias, and statistical variance of decision tree algorithms Technical Report. Corvallis, Oregon, Department of Computer Science, Oregon State University.

DIETTERICH, T. (1997) Fundamental experimental research in machine learning. Available at: http://web.engr.oregonstate.edu/~tgd/experimental-research/index.html (Cited: 27 October, 2009).

DIETTERICH, T. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation,* 10, 1895-1923.

DOHERTY, K. A. J., ADAMS, R. G. & DAVEY, N. (2007) Unsupervised learning with normalised data and non-Euclidean norms. *Applied soft computing,* 7(1). 203-217.

DOMINGO, C., GALVADA, R. & WATANABE, O. (2002) Adaptive sampling methods for scaling up knowledge discovery algorithms. *Data Mining and Knowledge Discovery,* 6, 131-152.

DOMINGOS, P. (2000a) A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning ,* 231-238. CA:Morgan Kaufmann.

DOMINGOS, P. (2000b) Bayesian averaging of classifiers and the overfitting problem. In *Proceedings of the Seventeenth International Conference on Machine Learning,* 223-230. CA:Morgan Kaufmann.

DOMINGOS, P. (2001) When and how to subsample: Report on the KDD-2001 Panel. *SIGKDD Explorations,* 3(2), 74-75.

ENGELBRECHT, A. P. (2002) *Computational intelligence: An introduction,* West Sussex:John Wiley & Sons.

FAN, C., MULLER, M. & REZUCHA, I. (1962) Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association* 57, 387-402.

FAN, W., LEE, W., STOLFO, J. & MILLER, M. (2000) A multiple model cost-sensitive approach for intrusion detection. *Lecture Notes in Computer Science*. Springer.

FAWCETT, T. (2001) Using rulesets to maximise ROC performance. *Proceedings of the IEEE International Conference on Data Mining (ICDM-2001),* 131-138.

FAWCETT, T. (2004) ROC graphs: Notes and practical considerations for researchers. HP Laboratories. Available from: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf (Cited: 1 March, 2010).

FAWCETT, T. (2006) An introduction to ROC analysis. *Pattern recognition Letters,* 27, 861-874.

FAYYAD, U., HAUSSLER, D. & STOLORZ, P. (1996) Mining Scientific Data *Communications of the ACM ,* 39 (11), 51-57.

FREUND, Y. & SCHAPIRE, R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences,* 55 (1), 119-139

FREY, L. J. & FISHER, D. H. (1999) Modeling Decision Tree Performance with the Power Law. IN HECKERMAN, D. & WHITTAKER, J. (Eds.) *Proceeding of the Seventh International Workshop on Artificial Intelligence and Statistics.* San Francisco, CA: Morgan Kauffman.

FRIEDMAN, J. (1997) On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery,* 1 (1), 55-77.

FU, Z., GOLDEN, B., LELE, S. & RAGHAVAN, S. (2003) Genetically engineered decision trees: population diversity produces smarter trees. *Operations Research,* 51 (6), 894-907.

FU, Z., GOLDEN, B.L., LELE, S., RAGHAVAN, S. & WASIL, E. (2006) Diversification for better decision trees. *Computers and Operations Research,* 51 (6), 894-907.

GEMAN, S., BIENENSTOCK, E. & DOURSAT, R. (1992) Neural networks and the bias/variance dilemma. *Neural computation,* 4, 1-58.

GIUDICI, P. (2003) *Applied Data Mining: Statistical Methods for Business and Industry,* Chichester:John Wiley & Sons.

GIUDICI, P. & FIGINI, s. (2009) *Applied Data Mining for Business and Industry, second edition,* Chichester:John Wiley & Sons.

GUYON, I. & ELISSEEFF, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* 3 1157-1182.

HALL, L. O., BOWYER, K. W., KEGELMEYER, P., MOORE, T. E. & CHAO, C. (2000) Distributed learning on very large data sets. *Proceedings of the Sixth International ACM SIGKDD.*

HALL, M. A. (1999) Correlation-based Feature Selection for Machine Learning, PhD Thesis. *Department of Computer Science* Hamilton, New Zealand, University of Waikato.

HALL, M. A. (2000) Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning* San Francisco, CA, Morgan Kaufmann.

HALL, M. A. & HOLMES, G. (2003) Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering* 15 (6), 1437- 1447.

HAND, D. J. (1997) *Construction and Assessment of Classification Rules,* Chichester:John Wiley & Sons

HAND, D. J. (1998) Data mining: statistics and more? *The American Statistician,* 52 (2), 112-118.

HAND, D. J. (1999) Statistics and data mining: intersection disciplines. *SIGKDD Explorations,* 1 (1), 16-19.

HAND, D. J., MANILA, H. & SMYTH, P. (2001) *Principles of Data Mining,* Cambridge, MA:MIT Press.

HAND, D. J. & TILL, R.J. (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning,* 45, 171-186.

HANLEY, J.A. & MCNEIL, B.J. (1982) The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology,* 143, 29-36.

HANSEN, L. K. & SALAMON, P. (1990) Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 12 (10), 993-1001.

HETTICH, S. & BAY, S. D. (1999) The UCI KDD archive [http://kdd.ics.uci.edu]. *Department of Information and Computer Science.* Irvine, CA, University of California.

HEVNER, A. R., MARCH, S. T., PARK, J. & RAM, S. (2004) Design science in information systems research. *MIS Quarterly,* 28 (1). 75-105.

HO, T. (1998) The random subspace method for constructing decision forests. *IEEE Trans. Pattern Analysis and Machine Intelligence,* 20 (8), 832-844.

HOEFFDING, W. (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association,* 58, 13-30.

IBA, W., WOGULIS, J. & LANGLEY, P. (1988) Trading off simplicity and coverage in incremental concept learning. IN ARBOR, A. (Ed.) *Proceedings of the 5th International Conference on Machine Learning.* Michigan:Morgan Kaufmann.

JAMES, G.M. (2003) Variance and bias for general loss functions. *Machine Learning,* 51,15-135.

JOHN, G. H. & LANGLEY, P. (1996) Static versus dynamic sampling for data mining. *Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining.* AAAI/MIT Press.

JONES, T. (1962) A note on sampling from tape files. *Communications of the ACM,* 5 (6). 343.

KANJI, G. (1999) *100 Statistical Tests,* London:Sage Publications.

KIM, E., KIM, W. & LEE, Y. (2002) Combination of multiple classifiers for the customer purchase behaviour prediction. *Decision Support Systems,* 34 (2), 167-175.

KINIVEN, J. & MANNILA, H. (1993) The power of sampling in knowledge discovery Technical Report C-1993-66. University of Helsinki.

KOHAVI, R., & JOHN, G.H. (1997) Wrappers for feature subset selection. *Artificial Intelligence,* 97, 273-324.

KOHAVI, R., MASON, R. J. & ZHENG, Z. (2004) Lessons and Challenges from Mining retail e-commerce data. *Machine Learning,* 57 83-113.

KOHAVI, R. & PROVOST, F. (1998) Glossary of terms. Special issue on applications of machine learning and the Knowledge Discovery process. *Machine Learning,* 30 (2). 271-274.

KOHAVI, R. & WOLPERT, D.H. (1996) Bias plus variance decomposition for zero-one loss functions. IN SAITTA, L. (Ed.) *Machine Learning: Proceedings of the Thirteenth International Conference,* 275-283. Morgan Kaufmann.

KONG, E. & DIETTERICH, T. (1995) Error-Correcting Output Coding Corrects Bias and Variance *Proceedings of the Twelfth International Conference on Machine Learning.* Morgan Kaufmann.

KROGH, A. & VEDELSBY, J. (1995) Neural network ensembles, cross validation and active learning. IN TESAURO, G., TOURETZKY, D. S. & LEEN, T. K. (Eds.) *Advances in Neural Information Processing Systems.* Cambridge MA: MIT Press.

KUBAT, M. & MATWIN, S. (1997) Addressing the curse of imbalanced data sets: One-sided selection. *Proceeding of the Fourteenth International Conference on Machine Learning.* San Francisco, CA, Morgan Kauffman.

KWOK, S. W. & CARTER, C. (1990) Multiple decision trees. *Uncertainty in Artificial Intelligence,* 4, 327-335.

LANGLEY, P., IBA, W. & THOMPSON, K. (1992) An analysis of Bayesian classifiers. *Proceedings, Tenth National Conference on Artificial Intelligence.* Menlo Park, CA, AAAI Press.

LASKOV, P., DÜSSEL, P., SCHÄFER, C. & RIECK, K. (2005) Learning intrusion detection: supervised or unsupervised? *ICAP: international conference on image analysis and processing.* Cagliari, Italy.

LEE, W., FAN, W., MILLER, M., STOLFO, S. & ZADOK, E. (2002) Toward cost-sensitive modeling for intrusion detection and response. *Journal of Computer Security,* 10 (1), 5-22.

LEE, W. & STOLFO, J. (2000) A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security,* 3 (4), 227-261.

LEUNG, K. & LECKIE, C. (2005) Unsupervised anomaly detection in network intrusion detection using clusters. IN ESTIVILL-CASTRO, V. (Ed.) *Proceedings of the Twenty-eighth Australasian conference on Computer Science* Newcastle, Australia, Australian Computer Society.

LINDEN, G., SMITH, B. & YORK, J. (2003) Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing,* 4(1).

LIU, H. & MOTODA, H. (1998) *Feature Selection for Knowledge Discovery and Data Mining.* Boston:Kluwer Academic Publishers.

LIU, H. & SETIONO, R. (1998a) Scalable feature selection for large sized databases. *Proceedings of the Fourth World Congress on Expert Systems (WCES'98)*. Morgan Kaufmann Publishers.

LIU, H. & SETIONO, R. (1998b) Some issues on scalable feature selection. *Expert Systems with Applications*, 15, 333-339.

LUGER, G. & STUBBLEFIELD, W. A. (1993) *Artificial Intelligence - Structures and Strategies for Complex Problem Solving, second edition.* CA:Benjamin Cummings Publishing Company.

LUTU, P. E. N. & ENGELBRECHT, A. P. (2006) A Comparative Study of Sample Selection methods for Classification. *South African Computer Journal,* 36, 69-85.

LUTU, P. E. N. & ENGELBRECHT, A. P. (2010) A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications,* 37 (1), 602-609.

MANSON, N. J. (2006) Is operations research really research? *Orion,* 22 (2), 155-180.

MARCH, S. T. & SMITH, G. F. (1995) Design and natural science research on information technology. *Decision Support Systems,* 15, 251-266.

MARTÍNEZ-MUÑOZ, G., HERNÁNDEZ-LOBATO, D. & SUÁREZ, A. (2009) An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), 245-259.

MITCHELL, T. M. (1997) *Machine Learning,* Burr Ridge, IL:WCB/McGraw-Hill.

MONTGOMERY, D. C., RONGER, G. C. & HUBELE, N. F. (2004) *Engineering Statistics,* New York:Wiley.

MOORE, A.W. & LEE, M.S. (1994)) Efficient algorithms for minimising cross validation error. In *Proceedings of the Eleventh International Conference on Machine Learning.* 190-198. New Brunswick, NJ: Morgan Kaufmann.

NEAGU, D., GUO, G. & WANG, S. (2006) An Effective Combination Based on Class-Wise Expertise of Diverse Classifiers for Predictive Toxicology Data Mining IN LI, X., ZAIANE, O. R. & LI, Z. (Eds.) *Advanced Data Mining and Applications.* Berlin, Springer Berlin / Heidelberg.

NEWELL, A. & SIMON, H. (1976) Computer science as empirical enemy: symbols and search. *Communication of the ACM,* 19 (2), 113-126.

NGWENYAMA, O. (2007) The seven basic claims of scientific research: an approach to analysing the structure of scientific argumentation in IS research papers. Working paper #iitm-2007-SR-187, Ryerson University, Toronto, Canada.

NGWENYAMA, O. K. & OSEI-BRYSON, K.-M. (2010) Using data mining to support information systems research: an approach for abduction and evaluation of hypotheses. To appear.

OATES, B. J. (1984) *Researching Information Systems and Computing.* London:SAGE Publications.

OLAFSSON, S., LI, X. & WU, S. (2008) Operations Research and data mining. *European Journal of Operations Research,* 19 (2) 113-126.

OLKEN, F. (1993) Random Sampling from Databases, PhD Thesis. *Department of Computer Science,* Berkeley. University of California at Berkeley.

OLKEN, F. & ROTEM, D. (1995) Random sampling from databases - A survey. (invited paper). *Statistics and Computing,* 5 (1), 25-42.

OOI, C. H., CHETTY, M. & TENG, S. W. (2007) Differential prioritization in feature selection and classifier aggregation for multiclass microarray datasets. *Data Mining and Knowledge Discovery,* 14, 329-366.

OSEI-BRYSON, K.-M. (2004) Evaluation of decision trees: a multi-criteria approach. *Computers and Operations Research,* 31 (11), 1933-1945.

OSEI-BRYSON, K.-M. (2007) Post-pruning in decision tree induction using multiple performance measures. *Computers and Operations Research,* 34, 3331-3345.

OSEI-BRYSON, K.-M. (2008) Post-pruning in regression tree induction: an integrated approach. *Expert Systems with Applications,* 34 (2), 1481-1490.

OSEI-BRYSON, K.-M., KAH, M. O. & KAH, J. M. L. (2008) Selecting predictive models for inclusion in an ensemble. *The 18th Triennial Conference of the International Federation of Operational Research Societies (IFORS 2008).* Sandton, Johannesburg, July 2008.

OSEI-BRYSON, K.-M. (2010) Towards supporting expert evaluation of clustering results using a data mining process model. *Information Sciences,* 180, 414-431.

PALMER, C. R. & FALOUTSOS, C. (2000) Density biased sampling: an improved method for data mining and clustering. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.* Dallas, Texas United States, ACM.

PEARL, J. (1984) *Heuristics: Intelligent Strategies for Computer Problem Solving,* Reading, MA:Addison-Wesley.

PHUA, C., LEE, V., SMITH, K. & GAYLER, R. (2005) A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review.* (SCI).

PROVOST, F., JENSEN, D. & OATES, T. (1999) Efficient progressive sampling. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* San Diego, CA, ACM.

PROVOST, F. & DOMINGOS, P. (2001) Well trained PETS: improving probability estimation trees. Working paper #IS-00-04, Stern School of Business, New York University, New York, NY 10012.

PROVOST, F. & FAWCETT, T. (2001) Robust classification for imprecise environments. *Machine Learning,* 42, 203-231.

QUINLAN, J. R. (1986) Induction of decision trees. *Machine Learning,* 1 81-106.

QUINLAN, J. R. (1993) *C4.5: Programs for Machine Learning,* San Francisco, CA:Morgan Kauffman.

QUINLAN, J. R. (2004) An Informal Tutorial, Rulequest Research. URL: http://www.rulequest.com (accessed: 28 October, 2005).

RAO, P. S. R. S. (2000) *Sampling Methodologies with Applications,* CRC, Florida:Chapman & Hall.

RIFKIN, R. & KLAUTAU, A. (2004) In defense of one-vs-all classification. *The Journal of Machine Learning Research,* 5, 101-141.

RYGIELSKI, C., WANG, J.-C. & YEN, D. C. (2002) Data Mining techniques for customer relationship management. *Technology in Society,* 24, 483-502.

SAMOILENKO, S. & OSEI-BRYSON, K.-M. (2008) Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees. *Expert Systems with Applications,* 34, 1568-1581.

SCHAFFER, C. (1994) A conservation law for generalisation performance. *Proceedings of the Eleventh Conference on Machine Learning,* 259-265, CA: Morgan-Kaufmann.

SCHAPIRE, R. (2003) The boosting approach to machine learning: An overview. *MSRI Workshop on Nonlinear Estimation and Classification.* Springer.

SCHLIMMER, J. S. (1987) Concept acquisition through representational adjustment (Technical Report 87-19). Doctoral dissertation. *Department of Information and Computer Science.* Irvine, University of California.

SHANON, C. E. & WEAVER, W. (1962) *The Mathematical Theory of Communication,* Urbana:University of Illinois Press.

SHEARER, C. (2000) The CRISP-DM model: the new blue print for data mining. *Journal of Data Warehousing,* 5(4), 13-22.

SHIN, S. W. & LEE, C. H. (2006) Using Attack-Specific Feature Subsets for Network Intrusion Detection. *Proceedings of the 19th Australian Conference on Artificial Intelligence.* Hobart, Australia.

SIMON, H. A. (1996) *The Science of the Artificial, 3rd Edition.* Cambridge, MA:MIT Press.

SMYTH, P. (2001) Data Mining at the interface of computer science and statistics. IN GROSSMAN, R. L., KAMATH, C., KEGELMEYER, P., KUMAR, V. & NAMBURU, R. R. (Eds.) *Data Mining for Scientific and Engineering Applications.* Dordrech, Netherlands, Kluwer Academic Publishers.

STIRLING, W. D. (2008) CAST - Computer Assisted Statistics Teaching. *Massey University, NZ.*

STOLFO, S. J., FAN, W., LEE, W., PRODROMIDIS, A. & CHAN, P. (2000) Cost-based modeling for fraud and intrusion detection: results from the JAM project. *DARPA Information Survivability Conference and Exposition.* Hilton Head, SC, USA.

STOLORZ, P. & DEAN, C. (1996) QuakeFinder: A scalable data mining system for detecting earthquake from the space. *Proceedings of the Second International Conference on Data Mining KDD-96.* Portland, Oregon, AAAI.

STOPPIGLIA, H., DREYFUS, G., DUBOIS, R. & OUSSAR, Y. (2003) Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research,* 3, 1399-1414.

SUN, J. & LI, H. (2008) Financial distress prediction based on serial combination of multiple classifiers. *Expert Systems With Applications,* 35 (5), 818-827.

THEUSINGER, C. & HUBER, K. P. (2000) Analysing the footsteps of your customer. *Proceedings of WEBKDD-2000.*

THOMAS, D. B., LUK, W., P.H.W., L. & J.D., V. (2007) Gaussian random number generators. *ACM Computer Survey,* 39 (4), 11:1-11:38.

TOIVONEN, H. (1996) Sampling large databases for association rules. *Proceedings of the Twenty-second Conference on Very Large Databases – VLDDB96.* Mumbai India, Morgan Kaufmann Publishers.

TOULMIN, S. E. (1958) *The Uses of Argumentation.* Cambridge, United Kingdom:Cambridge University Press.

TOULMIN, S., RIEKE, R. & JANIK, A. (1979) *An Introduction to Reasoning.* New York:Macmillan Publishing Co.

VAISHNAVI, V. & KUECHLER, W. (2004/5) Design Research in Information Systems. URL: http://desrist.org/design-research-in-information-systems (accessed 27 October, 2009).

VALIANT, L. G. (1984) A theory of the learnable. *Communications of the ACM,* 27 (11), 1134-1142.

VAN DER PUTTEN, P. & VAN SOMEREN, M. (2004) A bias-variance analysis of a real world learning problem: the COIL challenge 2000. *Machine Learning,* 57, 177-195.

VAPNIK, V. N. & CHERVONENKIS, A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications,* 16, 264-280.

VUK, M. & CURK, T. (2006) ROC curve, lift chart and calibration plot. *Metodološki Zvezki,* 3, 89-108.

WATANABE, O. (2005) Sequential sampling techniques for algorithmic learning theory. *Theoretical Computer Science,* 348, 3-14.

WAUGH, S. (1995) Extending and Benchmarking Cascade-Correlation, PhD thesis. *Computer Science Department.* Hobart, Tasmania, University of Tasmania.

WILCOX, R. R. (2001) *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy,* New York:Springer-Verlag.

WITTEN, H. I. & FRANK, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques, second edition,* San Francisco:Morgan Kaufmann.

WOLPERT, D. H. (1996) The lack of a priori distinctions between learning algorithms. *Neural Computation,* 8(7) 1341-1390.

WOLPERT, D. H. & Macready, G. (1997) No free lunch theorems for optimisation. *IEEE Transactions on Evolutionary Computation,* 1(1), 67-82.

WU, X., KUMAR, V., QUINLAN, J.R., GHOSH, J., YANG, Q., MOTODA, H., McLACHLAN, G.J., NG, A., LIU, B., YU, P.S., ZHOU, Z.-H., STEINBACH, M., HAND, D.J. & STEINBERG, D. (2008) Top 10 algorithms in data mining. *Knowledge Information Systems,* 14, 1-37.

YU, L. & LIU, H. (2004) Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research,* 5, 1205-1224.

Appendices

The appendices in this section provide definitions of the symbols used, general definitions of statistical measures and descriptive statistics for the datasets used in the experiments. Details of correlation measurements, details for the 5NN aggregation algorithm, and OVA and pVn model performance are provided. Information is also provided on suggestions on how to use commonly available statistical and database software to implement some of the steps for the proposed feature and training dataset selection methods. Finally, a list of publications and conference presentations arising from the research is given. The table below summarises the appendix contents.

Table of appendices

Appendix	Title	Description
A	Definition of symbols	Definition of symbols used in the thesis
B	Definitions of statistical measures	Definitions of statistical measures used in the thesis
C	Descriptive statistics for datasets	Descriptive statistics for forest cover type, KDD Cup 1999, Abalaone3C and mushroom
D	Correlation measurements	Details of correlation measurements and feature selection of chapter 4
E	Algorithm for 5NN aggregation	Details of algorithm for the combination of 5NN base model predictions
F	Predictive performance of OVA and pVn models	Detailed results for accuracy for single and aggregate models for chapters 7 and 8.
G	ROC analysis details	Computation of the AUC for one-versus-rest ROC analysis. Details of AUC computation results.
H	Using statistical and database software to implement dataset selection methods	Suggestions for using commonly available statistical and database software to implement dataset selection
I	Publications and Conference Presentations	Publications and conference presentations arising out of the work reported in this thesis

Appendix A

Definition of symbols

Table A.1: Symbols used in the thesis

Symbol	Meaning
$accuracy$	The predictive accuracy of a model
$B_i,..,B_v$	Binary features created through the process of binarisation of a qualitative feature with *v* levels
$corr(X,Y)$	The sample correlation coefficient between two random variables *X* and *Y*
$corr_{cf}(f)$	The sample correlation coefficient between a feature *f* and a class variable *C*
$corr_{ff}(f)$	The mean correlation between feature *f* and all other currently selected features
$c_1,\ldots,c_k$	The *k* levels of a class variable (number of classes for a prediction task)
C	A class variable for classification
$conf$	Probabilistic score assigned by a model to a class prediction as the level of confidence in the prediction
d	The number of predictive features (variables) that define the *d*-dimensional instance space for classification modeling
$1-\delta$	The probability of a learner being able to induce a hypothesis from data as in PAC
$error$	The prediction error of a model
$error_D, error_R$	Error difference and error ratio for measuring performance gains
$error_S, error_A$	Prediction errors of a single model and aggregate model respectively
ε	Prediction error as in PAC
E	Entropy function
f	A feature (predictor) used in predictive modeling
ϕ	The phi coefficient for measuring the level of association between two qualitative variables
g_i	A region of the instance space
G	The Gini concentration coefficient
h	A hypothesis as defined in machine learning
H	A set of hypotheses as defined in machine learning
H_0 and H_a	The null hypothesis and alternative hypotheses for statistical hypothesis testing
k	Number of classes for a classification problem
K	Number of folds for cross validation
$L_1,..,L_V$	Levels of a qualitative (nominal or ordinal) variable
λ	Cut-off score value for ROC analysis

Table A.1 continued	
Symbol	Meaning
m	A mapping or a function
M_A	General reference to a predictive model
μ_A	The population mean value of predictive accuracy of a model A
n	The size of a sample taken from a parent dataset
n_t	For sequential random sampling, n_t is the number of records already selected
N	The size of the parent dataset / database from which samples are taken
ova_i	The ith sub-problem for the prediction of class c_i in OVA classification
p	Probability of obtaining an experimental result given that the null hypothesis is true (p value)
P	Percentage value for a confidence interval (*P%* confidence interval)
P_r	Probability
PT	The number of partitions of a parent dataset
pred	Output of a predictive model
π_c and π_d	The probabilities of concordance and discordance used in the computation of Kendall's tau
r_{XY} , *r*	Pearson's sample correlation coefficient for two random variables *X* and *Y*
τ_{XY}	Kendall's sample correlation coefficient for two random variables X and Y
R^d	Super domain of real values for the random variables $X_1,..X_d$
RMsize, RQsize	For sequential random sampling, RMsize is the number of records still to be processed; RQsize is the number of records required for the sample
S_X	The sample standard deviation for random variable X
SU	Symmetrical uncertainty coefficient
σ_X	The population standard deviation for random variable X
si and spi	Situations for feature subset search
t	For sequential random sampling, t is the number of records processed so far
T	The *T* statistic for statistical hypothesis testing
u	Number of unselected features for heuristic feature subset search
v	Number of levels for a qualitative variable
V	Cramer's V statistic for measuring the level of association between two qualitative variables
VC(H)	The Vapnik-Chervonenkis dimension of a set of hypotheses *H* for a learning task
w	Number of features currently selected/processed by a feature selection method/algorithm
W	Number of candidate features for heuristic feature subset search

Table A1 continued

Symbol	Meaning
x *and* $x_1,..x_d$	A vector of predictive features (predictor variable) values (an instance)
x_q	A query (or test) instance to be classified / assigned a predicted value
X, Y	Random variables
Z	The Z statistic for statistical hypothesis testing
Z_P	Constant for the calculation of the *P%* confidence interval of the mean
$z\%$	Percentage of values to remove from each tail when winsorising variable values
Confusion matrix and ROC analysis symbols:	
Pos	Total number of positive instances
Neg	Total number of negative instances
TP	Number of positive instances predicted as positive
FN	Number of positive instances predicted as negative
TN	Number of negative instances predicted as negative
FP	Number of negative instances predicted as positive
$TPRATE$	Fraction of the positive instances predicted as positive
$FNRATE$	Fraction of the positive instances predicted as negative
$TNRATE$	Fraction of the negative instances predicted as negative
$FPRATE$	Fraction of the negative instances predicted as positive
$YRATE$	Fraction of test instances predicted as positive (used for lift analysis)

Appendix B

Definitions of statistical measures

A detailed discussion of the statistical measures used in this thesis is provided in this appendix. The entropy measure, Gini index of concentration, and measures of association (correlation) were used in the discussions of chapters 3, 4, 5 and 7.

B.1 Entropy definitions

The entropy function *E(X)* (Giudici, 2003; Shanon & Weaver, 1962) measures the amount of uncertainty, heterogeneity, information or randomness in the values of the qualitative or quantitative discrete random variable *X* and is defined as

$$E(X) = -\sum_{i} P_r(x_i)\log_2 P_r(x_i) \tag{B.1}$$

where $P_r(x_i)$ which is used as a shorthand notation for $P_r(X = L_i)$ is the probability that variable *X* has the value (level) L_i. The entropy of the random variable *X*, conditioned on the values of a second random variable *Y* is denoted as *E(X|Y)* and is defined as

$$E(X \mid Y) = -\sum_{j} P_r(y_j)\sum_{i} P_r(x_i \mid y_j)\log_2 P_r(x_i \mid y_j) \tag{B.2}$$

where $P_r(x_i \mid y_i)$ which is used as a shorthand for $P_r\big((X = L_i) \mid (Y = L_j)\big)$ is the conditional probability that random variable *X* has the value (level) L_i given that random variable *Y* has the value (level) L_j and is defined as

$$P_r(x_i \mid y_j) = \frac{P_r(x_i, y_j)}{P_r(y_j)} \tag{B.3}$$

where $P_r(x_i, y_j)$ is the probability of values x_i and y_j appearing together. The joint entropy of two random variables *X* and *Y* denoted as *E(X,Y)* is defined as

$$E(X,Y) = -\sum_{i} P_r(x_i, y_j)\log_2 P_r(x_i, y_j) \tag{B.4}$$

The difference between the entropy of *X, E(X)* and the entropy of *X* conditioned on *Y*, *E(X|Y)* is called the information gain *IG(X,Y)* and is defined as

$$IG(X,Y) = E(X) - E(X \mid Y) \tag{B.5}$$

$$IG(X,Y) = E(Y) - E(Y \mid X) \tag{B.6}$$

$$IG(X,Y) = E(X) + E(Y) - E(X,Y) \tag{B.7}$$

The information gain measures the amount of reduction in the entropy of *X* when the values of *X* are grouped based on the values of *Y*. As indicated by the equations (B.5) and (B.6), information gain *IG(X,Y)* is a symmetric measure from which the symmetrical uncertainty coefficient *SU* is derived. The *SU* coefficient is defined as

$$SU = 2.0x\left[\frac{IG(X,Y)}{E(X) + E(Y)}\right] \tag{B.8}$$

The SU coefficient was used for the experiments of chapters 5 and 7 as a measure of correlation (association) for qualitative features.

B.2 Measures of association

B.2.1 Pearson’s correlation coefficient

Pearson’s sample correlation coefficient, *r* (Wilcox, 2001), between two random variables *X* and *Y* is defined as

$$r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_X S_Y} \tag{B.9}$$

where S_X and S_Y are the standard deviations of *X* and *Y* respectively, and *n* is the sample size.

B.2.2 Kendall's correlation coefficient

Kendall's rank correlation coefficient *tau* (Wilcox, 2001) is defined as

$$\tau = \pi_c - \pi_d \tag{B.10}$$

where π_c and π_d are the probabilities of concordance and discordance respectively. A pair of observations, (x_1, y_1) and (x_2, y_2) shows concordance if $x_1 > x_2$ and $y_1 > y_2$ or $x_1 < x_2$ and $y_1 < y_2$, otherwise the pair shows discordance. The values π_c and π_d are computed for all possible pairs for a data sample. For a data sample of size *n*, there are $\frac{n(n-1)}{2}$ possible pairs. However, some pairs will be tied i.e. having neither concordance nor discordance.

B.2.3 Pearson's chi-square statistic

Pearson's chi-square statistic measures the level of association between two qualitative random variables *X* and *Y* (Giudici, 2003). The statistic is computed using the frequencies in a contingency table. A contingency table is a cross-tabulation which gives the frequencies of co-occurrence of the values (levels) of the variables *X* and *Y*. Pearson's chi-square statistic is defined as

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} \tag{B.11}$$

where *I* and *J* are respectively the number of rows and columns in the contingency table, n_{ij} are the observed frequencies in the cells of the contingency table and, n^*_{ij} are the expected frequencies for the cells of the contingency table under the null hypothesis of independence between *X* and *Y*.

The ϕ coefficient and Cramer's *V* coefficients are derived from Pearson's chi-square coefficient, and have the same interpretation as Pearson's *r* coefficient. The ϕ coefficient is defined as (Giudici, 2003)

$$\phi^2 = \frac{\chi^2}{n} \tag{B.12}$$

and Cramer's *V* coefficient is defined as

$$v^2 = \frac{\chi^2}{n.\min\{I-1, J-1\}} \tag{B.13}$$

The ϕ coefficient, Cramer's *V* coefficient, and symmetrical uncertainty coefficient can all be used to measure the level of association between two qualitative features.

B.3 Gini concentration coefficient

Suppose there are *n* entities on which a given property *EP* has been measured yielding *n* pairs of measurement values $\{(1, EP_1), \ldots, (i, EP_i), .., (n, EP_n)\}$ where *i* identifies the i^{th} entity and EP_i identifies the measurement value for the i^{th} entity. Let F_i be the cumulative percentage of the count of entities from the first to the i^{th} entity. Let Q_i be the cumulative percentage of the measurement values from the first measurement, EP_1 to the i^{th} measurement, EP_i. A summary statistic of the concentration of the measured property *EP* among the *n* entities is called the Gini concentration coefficient defined as

$$Gini = \frac{\sum_{i=1}^{n-1}(F_i - Q_i)}{\sum_{i=1}^{n-1} F_i} \tag{B.14}$$

The *Gini* measure equals 0 for minimum concentration and 1 for maximum concentration. Minimum concentration means that all *n* entities have equal values of the property *EP*. Maximum concentration means that only one entity possesses the property *EP* and all other *n-1* entities have a value of 0 for *EP*.

The Gini concentration coefficient is related to the Area Under the ROC curve (AUC) as follows: The EP property corresponds to the scores that are assigned by a probabilistic classifier. The AUC was discussed in section 4.7.

B.4 Computation of confidence intervals for the mean

A *P%* confidence interval for the mean is an interval that is expected with probability *P%* to contain the true value of the population mean (Mitchell, 1997). Laplace's estimate of the confidence interval of the population mean is defined as

$$CI = \left(\bar{x} - Z_P \frac{S_X}{\sqrt{n}}, \bar{x} + Z_P \frac{S_X}{\sqrt{n}} \right) \quad \text{(B.15)}$$

where $\bar{x}$ is the sample mean for random variable *X*, S_X is the sample standard deviation, and *n* is the sample size (Wilcox, 2001; Mitchell, 1997). Different values of Z_P are used to obtain different confidence intervals. A value of $Z_P = 1.96$ is used for the 95% confidence interval. A value of $Z_P = 2.58$ is used for the 99% confidence interval (Wilcox, 2001; Mitchell, 1997).

Appendix C

Descriptive statistics for the datasets

The descriptive statistics for the datasets used in the experiments are presented in this section.

C.1 Forest cover type dataset

Figure C.1 provides the class frequencies and a graphic representation for the forest cover type dataset classes. Tables C.1 and C.2 show the descriptive statistics for the qualitative and quantitative variables in the forest cover type dataset.

Class label	Class name	Percent
1	Spruce / fir	36.5
2	Lodgepole pine	48.8
3	Panderosa pine	6.2
4	Cottonwood / Willow	.5
5	Aspen	1.6
6	Douglas-fir	3.0
7	Krummholz	3.5

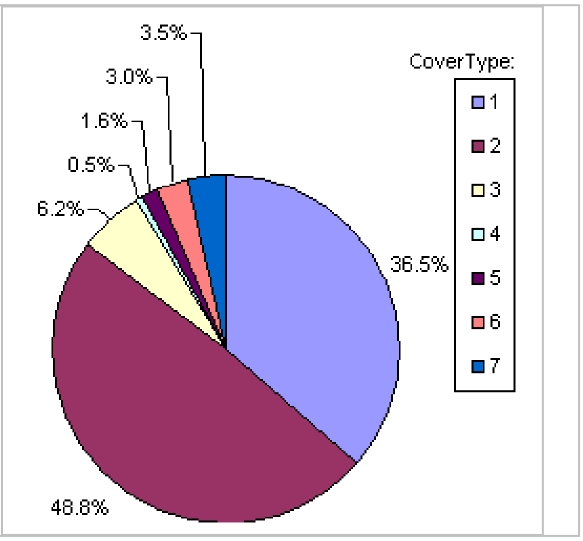


Figure C.1: Class frequencies for the forest cover type class variable (covertype)

Table C.1: Descriptive statistics for the quantitative variables in the forest cover type dataset

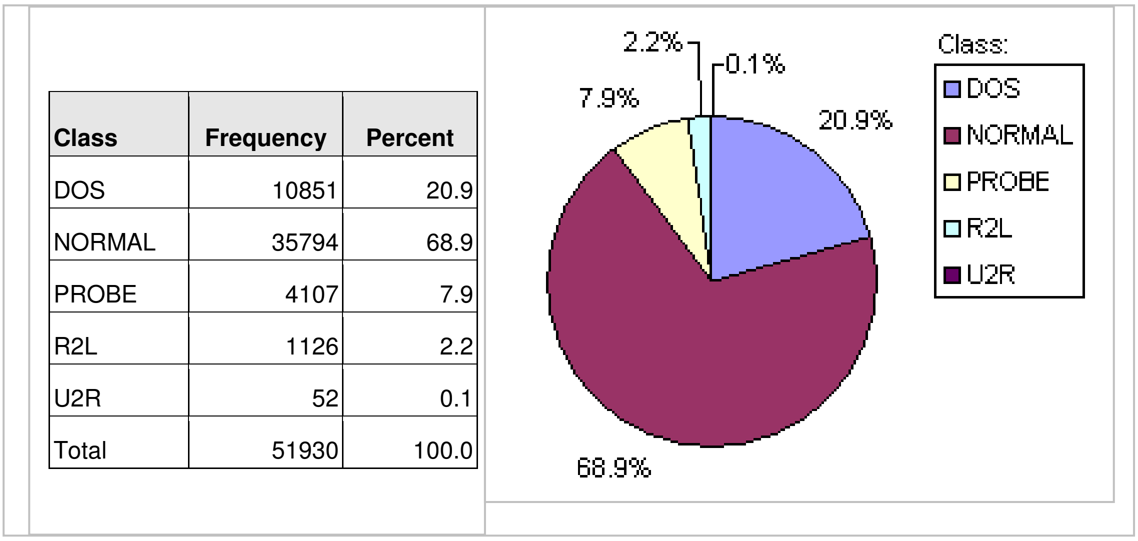
	Minimum	Maximum	Mean	Standard Deviation	Coefficient of variation (CV)
Aspect	0	360	155.7	111.9	0.7
Elevation	1859	3858	2959.4	280.0	0.1
Slope	0	66	14.1	7.5	0.5
HorizDistToHydro	0	1397	269.4	212.5	0.8
VertDistToHydro	-173	601	46.4	58.3	1.3
HorizDistToRoad	0	7117	2350.2	1559.3	0.7
HillShade9am	0	254	212.2	26.8	0.1
HillShadeNoon	0	254	223.3	19.8	0.1
HillShade3pm	0	254	142.5	38.3	0.3
HorizDistToFire	0	7173	1980.3	1324.2	0.7

Table C.2: Descriptive statistics for the qualitative variables for the forest cover type dataset

Variable name	Percentage for '0'	Percentage for '1'	Variable name	Percentage for '0'	Percentage for '1'
WildernessArea1	55.1	44.9	SoilType19	99.3	0.7
WildernessArea2	94.9	5.1	SoilType20	98.4	1.6
WildernessArea3	56.4	43.6	SoilType21	99.9	0.1
WildernessArea4	93.6	6.4	SoilType22	94.3	5.7
SoilType1	99.5	0.5	SoilType23	90.1	9.9
SoilType2	98.7	1.3	SoilType24	96.3	3.7
SoilType3	99.2	0.8	SoilType25	99.9	0.1
SoilType4	97.9	2.1	SoilType26	99.6	0.4
SoilType5	99.7	0.3	SoilType27	99.8	0.2
SoilType6	98.9	1.1	SoilType28	99.8	0.2
SoilType7	99.98	0.02	SoilType29	80.2	19.8
SoilType8	99.97	0.03	SoilType30	94.8	5.2
SoilType9	99.8	0.2	SoilType31	95.6	4.4
SoilType10	94.4	5.6	SoilType32	91	9
SoilType11	97.9	2.1	SoilType33	92.2	7.8
SoilType12	94.8	5.2	SoilType34	99.7	0.3
SoilType13	97	3	SoilType35	99.7	0.3
SoilType14	99.9	0.1	SoilType36	100	0
SoilType15	100	0	SoilType37	99.9	0.1
SoilType16	99.5	0.5	SoilType38	97.3	2.7
SoilType17	99.4	0.6	SoilType39	97.6	2.4
SoilType18	99.7	0.3	SoilType40	98.5	1.5

C.2 KDD Cup 1999 dataset

Figure C.2 provides the class frequencies and a graphic representation for the KDD Cup 1999 dataset classes. Tables C.3 and C.4 give the descriptive statistics for the variables in the KDD Cup 1999 dataset.



Class	Frequency	Percent
DOS	10851	20.9
NORMAL	35794	68.9
PROBE	4107	7.9
R2L	1126	2.2
U2R	52	0.1
Total	51930	100.0

Figure C.2: Class frequencies for the KDD Cup 1999 training dataset derived class variable (class)

Table C.3: Descriptive statistics for the quantitative variables for the KDD Cup 1999 training dataset

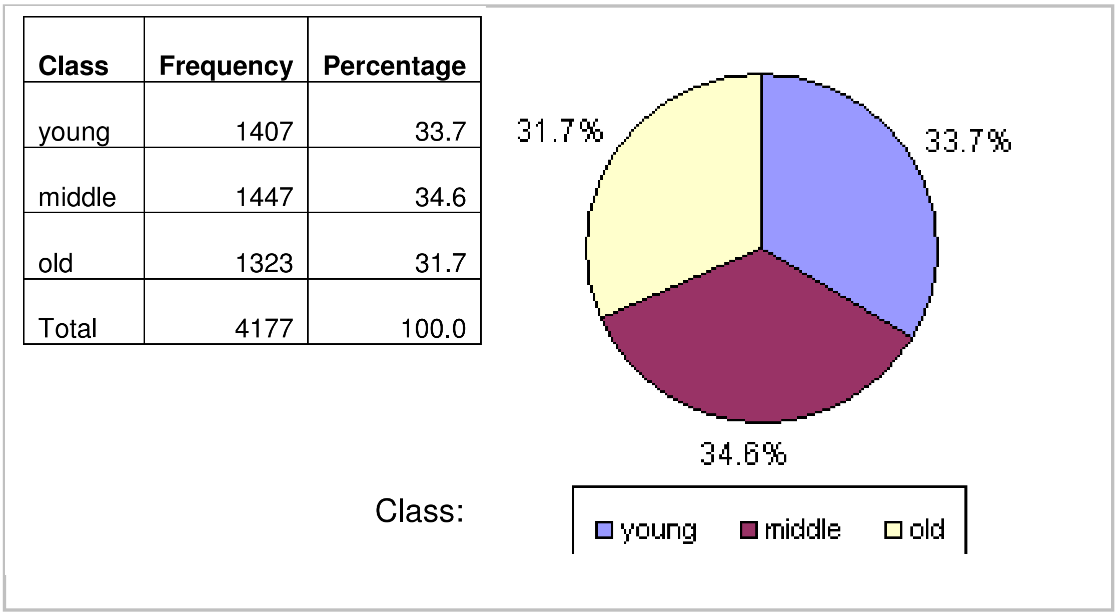
Variable name	Minimum	Maximum	Mean	Standard Deviation	Coefficient of variation (CV)
Counted	0	511	53.3	120.4	2.3
DiffSrvRate	0	1	0.1	0.2	2.0
DstBytes	0	5,155,468.00	3,758.50	99,612.90	26.5
DstHostCount	1	255	191	93.2	0.5
DstHostDiffSrvRate	0	1	0.2	0.3	1.5
DstHostRerrorRate	0	1	0.1	0.2	2.0
DstHostSameSrcPortRate	0	1	0.3	0.4	1.3
DstHostSameSrvRate	0	1	0.6	0.4	0.7
DstHostSerrorRate	0	1	0.1	0.3	3.0
DstHostSrvCount	1	255	120.9	107.3	0.9
DstHostSrvDiffHostRate	0	1	0	0.1	undefined
DstHostSrvRerrorRate	0	1	0.1	0.2	2.0
DstHostSrvSerrorRate	0	1	0.1	0.3	3.0
Duration	0	58,329.00	455.5	2,140.00	4.7
Hot	0	30	0.3	2.4	8.0
NumAccessFiles	0	8	0	0.1	undefined
NumCompromised	0	884	0.1	5.5	55.0
NumFailedLogins	0	5	0	0	undefined
NumFileCreations	0	28	0	0.3	undefined
NumOutboundCmds	0	0	0	0	undefined
NumRoot	0	993	0.1	6.2	62.0
NumShells	0	2	0	0	undefined
RerrorRate	0	1	0.1	0.2	2.0
RootShell	0	1	0	0	undefined
SameSrvRate	0	1	0.8	0.4	0.5
SerrorRate	0	1	0.1	0.3	3.0
SrcBytes	0	693,000,000.00	23,327.40	3,047,960.00	130.7
SrvCount	0	511	20	73.9	3.7
SrvDiffHostRate	0	1	0.1	0.3	3.0
SrvRerrorRate	0	1	0.1	0.3	3.0
SrvSerrorRate	0	1	0.1	0.3	3.0
SUAttempted	0	2	0	0	undefined
Urgent	0	3	0	0	undefined
WrongFragment	0	3	0.1	0.4	4.0

Table C.4: Descriptive statistics for the qualitative variables for the KDD Cup 1999 training dataset

Variable	Level description	Level names	Frequency%
ProtocolType	3 levels	icmp	7.3
		tcp	53.5
		udp	39.2
Service	64 levels	domain_u	11.3
		ftp_data	9.1
		http	14.3
		private	19.4
		smtp	9.9
		all other services	36
Flag	9 levels	SF	82.1
		S0	10.7
		all other flags	7.2
Land	2 levels	0	99.96
		1	0.04
LoggedIn	2 levels	0	67
		1	33
IsHostLogin	2 levels	0	100
		1	0
IsGuestLogin	2 levels	0	98.7
		1	1.3

C.3 Abalone3C dataset

Figure C.3 provides the class frequencies and graphic representation for the abalone3C dataset classes. Table C.5 gives the descriptive statistics for the variables.



Class	Frequency	Percentage
young	1407	33.7
middle	1447	34.6
old	1323	31.7
Total	4177	100.0

Figure C.3: Class frequencies for the abalone3C class variable (age)

Table C.5: Descriptive statistics for the quantitative variables of abalone3C

Variable	Minimum	Maximum	Mean	Standard Deviation	Coefficient of variation (CV)
Length	15.0	163.0	104.8	24.0	0.2
Diameter	11.0	130.0	81.6	19.8	0.2
Height	0.0	226.0	27.9	8.4	0.3
WholeWeight	0.4	565.1	165.7	98.1	0.6
ShuckledWeight	0.2	297.6	71.9	44.4	0.6
VisceraWeight	0.1	152.0	36.1	21.9	0.6
ShellWeight	0.3	201.0	47.8	27.8	0.6

The qualitative variable gender has three levels with absolute frequencies of: 1528 for male (M), 1307 for female (F) and 1342 for infant (I).

C.4 Wine quality datasets

Figure C.4 provides the class frequencies and graphic representation for the wine quality (white) dataset classes. The two minority classes: 3 (20 instances) and 9 (5 instances) were removed from the dataset. Table C.6 gives the descriptive statistics for the variables.

Class	Frequency	Percentage
4	163	3.3
5	1457	29.9
6	2198	45.1
7	880	18.1
8	175	3.6
Total	4873	100.0

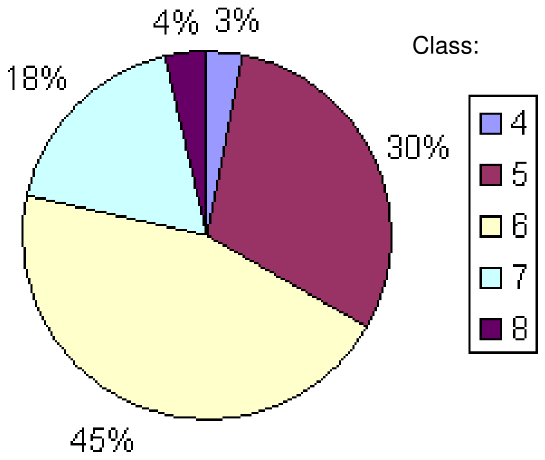


Figure C.4: Class frequencies for the wine quality (white) class variable (quality)

Table C.6 Descriptive statistics for the Wine quality (white) dataset variables

Variable	Minimum	Maximum	Mean	Standard Deviation	Coeff of variation (CV)
FixedAcidity	3.8	14.2	6.9	0.8	0.1
VolatileAcidity	0.1	1.1	0.3	0.1	0.4
CitricAcid	0.0	1.7	0.3	0.1	0.4
ResidualSugar	0.6	65.8	6.4	5.1	0.8
Chlorides	0.0	0.3	0.0	0.0	0.5
FreeSulfurDioxide	2.0	289.0	35.3	17.0	0.5
TotalSulfurDioxide	9.0	440.0	138.4	42.5	0.3
Density	1.0	1.0	1.0	0.0	0.0
pH	2.7	3.8	3.2	0.2	0.0
Sulphates	0.2	1.1	0.5	0.1	0.2
Alcohol	8.0	14.2	10.5	1.2	0.1

C.5 Mushroom dataset

Table C.7 gives the descriptive statistics for the mushroom dataset. The variables for this dataset are all qualitative nominal.

Table C.8 Descriptive statistics for the mushroom dataset variables

Variable	Level description	Level name	Frequ-ency%	Variable	Level description	Level name	Frequ-ency%
CapShape	6 levels	FLAT	39.1	StalkRoot	5 levels	EQUAL	16.3
		CONVEX	45.1			BULBOUS	45.2
		All other	15.8			UNKNOWN	29.5
CapSurface	4 levels	GROOVES	0.05			All other	9.0
		SMOOTH	31.9	StalkSfAbvRing	4 levels	SILKY	28.3
		FIBROUS	29.2			SMOOTH	63.2
		SCALY	38.8			All other	8.5
CapColor	10 levels	WHITE	12.4	StalkSfBlRing	4 levels	SILKY	27.4
		RED	17.8			SMOOTH	60.3
		YELLOW	12.7			All other	12.3
		BROWN	27.6	StalkClAbvRing	9 levels	WHITE	56.4
		GRAY	24.9			PINK	22.2
		All other	4.6			All other	21.4
Bruises?	2 levels	NO	59.9	StalkClBlRing	9 levels	WHITE	55.1
		BRUISES	40.1			PINK	22.2
Odor	9 levels	FOUL	25.7			All other	22.6
		NONE	45.2	VeilType	1 level	PARTIAL	100.0
		All other	29.1	VeilColor	4 levels	WHITE	97.6
GillAttach	2 levels	FREE	97.4			All other	2.4
		ATTACHED	2.6	RingNumber	3 levels	ONE	92.3
GillSpace	2 levels	CROWDED	18.9			All other	7.7
		CLOSE	81.1	RingType	5 levels	LARGE	15.4
GillSize	2 levels	NARROW	30.1			PENDANT	47.1
		BROAD	69.9			EVANESCENT	36.3
GillColor	12 levels	WHITE	14.6			All other	1.1
		PINK	18.5	SporePrintColor	9 levels	BLACK	23.8
		BUFF	20.5			WHITE	28.8
		BROWN	13.2			CHOCOLATE	19.4
		All other	33.1			BROWN	24.9
		GRAY	8.9			All other	3.1
StalkShape	2 levels	ENLARGING	42.2	Population	6 levels	SOLITARY	20.3
		TAPERING	57.8			SEVERAL	48.3
Habitat	7 levels	PATHS	13.6			SCATTERED	16.3
		LEAVES	10.2			All other	15.0
		GRASSES	28.6	**Class**	**2 levels**	**EDIBLE**	**53.3**
		WOODS	37.5			**POISONOUS**	**46.7**
		All other	10.1				

Appendix D

Correlation measurements for feature selection

The details of feature selection discussed in chapters 5 and 7 are presented in this appendix. Tables D.1 to D.4 show the class-feature correlations and the number of features selected by the t-test and the probes using Pearson's r and Kendall's tau measures of correlation for the forest cover type dataset.

Table D.1: Feature selection for Forest cover type

Sample size for correlation measurement	Selection criteria (Number of selected features)	Top 10 features				
		Feature	Mean $Corr_{cf}$	StDev	95% CI of mean	
					Low	High
100	**Pearson's r** t-test (3)	WildernessArea4	0.2	0.06	0.16	0.24
		SoilType38	0.14	0.04	0.12	0.16
		Elevation	0.14	0.04	0.12	0.16
500 and 1000	Pearson's r t-test (6) (WildernessArea1 is selected for sample size 500, SoilType10 is selected for size 1000)	WildernessArea4	0.22	0.02	0.21	0.23
		SoilType12	0.16	0.02	0.15	0.17
		SoilType22	0.14	0.03	0.12	0.16
		Elevation	0.13	0.02	0.12	0.14
		WildernessArea1	0.12	0.01	0.11	0.13
		SoilType38	0.12	0.02	0.11	0.13
100	**Kendall's tau:** t-test (20) Uniform probe (26) Uniform binary probe (21) Gaussian probe (31)	WildernessArea4	0.58	0.15	0.49	0.67
		SoilType12	0.51	0.19	0.39	0.63
		SoilType38	0.44	0.1	0.38	0.50
		SoilType22	0.43	0.17	0.32	0.54
		SoilType10	0.4	0.13	0.32	0.48
		SoilType39	0.38	0.17	0.27	0.49
		SoilType4	0.35	0.2	0.23	0.47
		SoilType23	0.35	0.15	0.26	0.44
		SoilType11	0.32	0.16	0.22	0.42
		SoilType30	0.31	0.19	0.19	0.43
500	**Kendall's tau** t-test (35) Uniform probe (47) Uniform binary probe (44) Gaussian probe (47)	WildernessArea4	0.81	0.03	0.79	0.83
		SoilType12	0.72	0.08	0.67	0.77
		SoilType38	0.6	0.08	0.55	0.65
		SoilType39	0.58	0.09	0.52	0.64
		SoilType2	0.58	0.15	0.49	0.67
		SoilType22	0.57	0.1	0.51	0.63
		SoilType4	0.57	0.12	0.50	0.64
		SoilType6	0.56	0.11	0.49	0.63
		SoilType13	0.56	0.11	0.49	0.63
		SoilType40	0.48	0.11	0.41	0.55
1000	**Kendall's tau:** t-test (38) Uniform probe (48) Uniform binary probe (47) Gaussian probe (49)	WildernessArea4	0.86	0.02	0.85	0.87
		SoilType12	0.7	0.07	0.66	0.74
		SoilType1	0.69	0.05	0.66	0.72
		SoilType38	0.68	0.08	0.63	0.73
		SoilType39	0.68	0.08	0.63	0.73
		SoilType2	0.64	0.1	0.58	0.70
		SoilType4	0.64	0.05	0.61	0.67
		SoilType6	0.6	0.1	0.54	0.66
		SoilType22	0.59	0.1	0.53	0.65
		SoilType10	0.58	0.05	0.55	0.61

Table D.2: Feature selection for forest cover type using Kendall's tau and a Gaussian probe

Rank	Feature	Kendall's tau		Feature 95% CI		Gaussian probe 95% CI		Select
		Mean	Stdev	Low	High	Low	High	
1	WildernessArea4	0.86	0.02	0.84	0.87	0.02	0.05	yes
2	SoilType12	0.70	0.07	0.66	0.75	0.02	0.05	yes
3	SoilType1	0.69	0.05	0.65	0.72	0.02	0.05	yes
4	SoilType38	0.68	0.08	0.63	0.73	0.02	0.05	yes
5	SoilType39	0.68	0.08	0.62	0.73	0.02	0.05	yes
6	SoilType2	0.64	0.10	0.58	0.70	0.02	0.05	yes
7	SoilType4	0.64	0.05	0.61	0.67	0.02	0.05	yes
8	SoilType6	0.60	0.10	0.54	0.67	0.02	0.05	yes
9	SoilType22	0.59	0.10	0.53	0.65	0.02	0.05	yes
10	SoilType10	0.58	0.05	0.55	0.61	0.02	0.05	yes
11	SoilType3	0.55	0.10	0.48	0.61	0.02	0.05	yes
12	SoilType40	0.55	0.10	0.49	0.61	0.02	0.05	yes
13	SoilType13	0.53	0.10	0.47	0.59	0.02	0.05	yes
14	SoilType11	0.48	0.08	0.43	0.52	0.02	0.05	yes
15	SoilType35	0.44	0.09	0.39	0.50	0.02	0.05	yes
16	SoilType18	0.44	0.17	0.34	0.54	0.02	0.05	yes
17	SoilType17	0.43	0.16	0.34	0.53	0.02	0.05	yes
18	SoilType26	0.43	0.16	0.33	0.53	0.02	0.05	yes
19	SoilType34	0.40	0.18	0.29	0.51	0.02	0.05	yes
20	SoilType23	0.40	0.04	0.37	0.43	0.02	0.05	yes
21	WildernessArea2	0.39	0.12	0.31	0.47	0.02	0.05	yes
22	SoilType5	0.36	0.22	0.22	0.50	0.02	0.05	yes
23	SoilType19	0.35	0.17	0.25	0.46	0.02	0.05	yes
24	SoilType30	0.34	0.10	0.28	0.40	0.02	0.05	yes
25	SoilType16	0.33	0.13	0.25	0.41	0.02	0.05	yes
26	SoilType21	0.32	0.20	0.20	0.44	0.02	0.05	yes
27	SoilType29	0.30	0.04	0.27	0.32	0.02	0.05	yes
28	WildernessArea1	0.28	0.03	0.27	0.30	0.02	0.05	yes
29	SoilType9	0.28	0.16	0.19	0.38	0.02	0.05	yes
30	Elevation	0.28	0.01	0.27	0.29	0.02	0.05	yes
31	SoilType24	0.26	0.09	0.20	0.32	0.02	0.05	yes
32	SoilType14	0.23	0.22	0.10	0.37	0.02	0.05	yes
33	SoilType31	0.22	0.08	0.17	0.27	0.02	0.05	yes
34	SoilType28	0.21	0.15	0.12	0.31	0.02	0.05	yes
35	SoilType32	0.21	0.02	0.19	0.22	0.02	0.05	yes
36	SoilType33	0.18	0.04	0.16	0.21	0.02	0.05	yes
37	SoilType8	0.18	0.16	0.08	0.27	0.02	0.05	yes
38	SoilType20	0.16	0.03	0.14	0.18	0.02	0.05	yes
39	HorizDistToRoad	0.16	0.01	0.15	0.17	0.02	0.05	yes
40	HorizDistToFire	0.16	0.01	0.15	0.16	0.02	0.05	yes
41	SoilType27	0.15	0.15	0.05	0.24	0.02	0.05	yes
42	Slope	0.12	0.02	0.11	0.14	0.02	0.05	yes
43	HillShade9am	0.08	0.02	0.07	0.10	0.02	0.05	yes
44	VertDistToHydro	0.07	0.02	0.06	0.08	0.02	0.05	yes
45	HorizDistToHydro	0.07	0.02	0.06	0.08	0.02	0.05	yes
46	WildernessArea3	0.07	0.03	0.05	0.09	0.02	0.05	yes
47	HillShadeNoon	0.07	0.02	0.06	0.08	0.02	0.05	yes
48	Aspect	0.05	0.02	0.03	0.06	0.02	0.05	yes
49	HillShade3pm	0.04	0.02	0.03	0.06	0.02	0.05	yes
50	**Probe1GaussCont**	**0.04**	**0.02**	**0.02**	**0.05**	**0.02**	**0.05**	**no**

Tables D.4 and D.5 show the class-feature correlations using Pearson's r, Kendall's tau and SU coefficient, and the number of features selected by the t-test, probes and decision rule-based algorithm for the KDDCup 1999 dataset.

Table D.3 Features selected by the decision rule-based search algorithm for different inputs

Input feature set selected by:	Number of selected features		Top 10 features for all methods	
			Feature	mean corr_{cf}
No pre-selection (54 features + 3 probes)	42		WildernessArea4	0.855
			SoilType2	0.642
Gaussian probe (49 features)	41		SoilType40	0.547
			SoilType38	0.676
Uniform probe (48 features)	41		SoilType4	0.638
			SoilType1	0.686
Uniform binary probe (47 features)	41		SoilType3	0.548
			SoilType6	0.603
t-test for means (36 features)	36		SoilType13	0.527
			SoilType39	0.676

Table D.4: Feature selection for KDD Cup 1999

Sample size for correlation measurement	Selection criteria (Number of selected features)	Top 10 features				
					95% CI of mean	
		Feature	Mean Corr_{cf}	StDev	Low	High
1000	Pearson's r: t-test (21) Uniform probe (32) Uniform binary probe (31) Gaussian probe (31)	SameSrvRate	0.53	0.02	0.52	0.54
		SerrorRate	0.51	0.02	0.50	0.52
		DstHostSerrorRate	0.51	0.02	0.50	0.52
		Counted	0.51	0.02	0.50	0.52
		SrvSerrorRate	0.50	0.02	0.49	0.51
		DstHostSrvSerrorRate	0.50	0.02	0.49	0.51
		Flag	0.43	0.02	0.42	0.44
		DstHostRerrorRate	0.36	0.03	0.34	0.38
		SrvRerrorRate	0.35	0.03	0.33	0.37
		RerrorRate	0.34	0.03	0.32	0.36
500	Kendall's tau: t-test (34) Uniform probe (36) Uniform binary probe (36) Gaussian probe (36)	SrvSerrorRate	0.90	0.02	0.89	0.91
		SerrorRate	0.87	0.02	0.86	0.88
		NumCompromised	0.85	0.03	0.83	0.87
		DstHostSrvSerrorRate	0.83	0.04	0.81	0.85
		WrongFragment	0.81	0.04	0.78	0.84
		DstHostSerrorRate	0.81	0.02	0.80	0.82
		SrvRerrorRate	0.80	0.04	0.78	0.82
		Hot	0.78	0.04	0.76	0.80
		DstHostSrvRerrorRate	0.76	0.05	0.73	0.79
		RerrorRate	0.76	0.05	0.73	0.79
1000	Kendall's tau: t-test (30) Uniform probe (36) Uniform binary probe (35) Gaussian probe (36)	SerrorRate	0.92	0.01	0.91	0.93
		NumCompromised	0.92	0.03	0.90	0.94
		SrvSerrorRate	0.91	0.01	0.90	0.92
		WrongFragment	0.9	0.01	0.89	0.91
		DstHostSrvSerrorRate	0.85	0.01	0.84	0.86
		DstHostSrvRerrorRate	0.85	0.01	0.84	0.86
		SrvRerrorRate	0.85	0.02	0.84	0.86
		Hot	0.84	0.03	0.82	0.86
		DstHostSerrorRate	0.84	0.02	0.83	0.85
		RerrorRate	0.82	0.03	0.80	0.84

Table D.5: Feature selection for KDD Cup1999 using Kendall's tau and the Gaussian probe

Rank	Feature	**Mean**	**StDev**	Feature 95% CI		Gauss probe 95% CI		Select
				Low	**High**	**Low**	**High**	
1	SerrorRate	0.92	0.01	0.91	0.92	0.02	0.04	yes
2	NumCompromised	0.92	0.03	0.90	0.93	0.02	0.04	yes
3	SrvSerrorRate	0.91	0.01	0.91	0.92	0.02	0.04	yes
4	WrongFragment	0.90	0.01	0.89	0.91	0.02	0.04	yes
5	DstHostSrvSerrorRate	0.85	0.01	0.85	0.86	0.02	0.04	yes
6	DstHostSrvRerrorRate	0.85	0.01	0.84	0.85	0.02	0.04	yes
7	SrvRerrorRate	0.85	0.02	0.83	0.86	0.02	0.04	yes
8	Hot	0.84	0.03	0.83	0.86	0.02	0.04	yes
9	DstHostSerrorRate	0.84	0.02	0.82	0.85	0.02	0.04	yes
10	RerrorRate	0.82	0.03	0.80	0.84	0.02	0.04	yes
11	SameSrvRate	0.82	0.01	0.81	0.83	0.02	0.04	yes
12	DstHostRerrorRate	0.80	0.03	0.79	0.82	0.02	0.04	yes
13	DiffSrvRate	0.73	0.02	0.71	0.74	0.02	0.04	yes
14	NumRoot	0.68	0.10	0.62	0.74	0.02	0.04	yes
15	Counted	0.63	0.01	0.62	0.64	0.02	0.04	yes
16	DstBytes	0.58	0.06	0.55	0.62	0.02	0.04	yes
17	SrcBytes	0.49	0.05	0.46	0.52	0.02	0.04	yes
18	SrvDiffHostRate	0.46	0.08	0.41	0.50	0.02	0.04	yes
19	DstHostSrvDiffHostRate	0.44	0.05	0.41	0.47	0.02	0.04	yes
20	Flag	0.43	0.02	0.41	0.44	0.02	0.04	yes
21	SrvCount	0.42	0.02	0.41	0.44	0.02	0.04	yes
22	DstHostCount	0.37	0.03	0.35	0.39	0.02	0.04	yes
23	DstHostSrvCount	0.31	0.04	0.29	0.34	0.02	0.04	yes
24	NumFailedLogins	0.30	0.23	0.16	0.44	0.02	0.04	yes
25	NumFileCreations	0.30	0.08	0.25	0.35	0.02	0.04	yes
26	DstHostSameSrcPortRate	0.28	0.05	0.25	0.31	0.02	0.04	yes
27	Duration	0.25	0.02	0.24	0.27	0.02	0.04	yes
28	Service	0.24	0.01	0.23	0.24	0.02	0.04	yes
29	DstHostSameSrvRate	0.22	0.04	0.20	0.25	0.02	0.04	yes
30	NumShells	0.20	0.16	0.11	0.30	0.02	0.04	yes
31	NumAccessFiles	0.18	0.20	0.06	0.30	0.02	0.04	yes
32	ProtocolType	0.15	0.02	0.14	0.16	0.02	0.04	yes
33	DstHostDiffSrvRate	0.14	0.04	0.12	0.17	0.02	0.04	yes
34	RootShell	0.11	0.15	0.02	0.20	0.02	0.04	no
35	LoggedIn	0.08	0.01	0.08	0.09	0.02	0.04	yes
36	IsGuestLogin	0.04	0.01	0.03	0.05	0.02	0.04	yes
37	Urgent	0.03	0.11	-0.03	0.10	0.02	0.04	no
38	**Probe1GaussCont**	**0.03**	**0.02**	**0.02**	**0.04**	**0.02**	**0.04**	**no**

Tables D.7 and D.9 show the class-feature correlations using Pearson's r, Kendall's tau and the SU coefficient, and the number of features selected by the t-test, probes and decision rule-based algorithm for the abalone3C and mushroom datasets. Table D.8 shows the feature-feature correlations for abalone3C.

Table D.6: KDD Cup 1999 feature selection by decision rule

Input feature set selected by:	Number of selected features		Top 10 for no-preselection (32 features selected)	
			Feature	mean corr$_{cf}$
No pre-selection (41 features + 3 probes)	32		SerrorRate	0.92
			DstHostRerrorRate	0.81
Gaussian probe (36 features)	34		NumRoot	0.68
			WrongFragment	0.90
Uniform probe (36 features)	34		Flag	0.43
			NumFailedLogins	0.30
Uniform binary probe (35 features)	34		DstHostSerrorRate	0.84
			DstHostSrvCount	0.31
t-test for means (30 features)	30		SrvCount	0.42
			DstHostCount	0.37

Table D.7: Feature selection for Abalone using Pearson's r and Kendall's tau

Sample size	Selection criteria (Number of selected features)	Selected features				
		Feature	Mean Corr$_{CF}$	StDev	95% CI of mean	
					Low	High
500 and 1000	Pearson's r: t-test (5) probes do not eliminate any features	Diameter	0.41	0.02	0.40	0.42
		ShellWeight	0.4	0.02	0.39	0.41
		WholeWeight	0.38	0.02	0.37	0.39
		VisceraWeight	0.38	0.02	0.37	0.39
		ShuckledWeight	0.34	0.02	0.33	0.35
500	Kendall's tau: t-test (6) probes do not eliminate any features	Height	0.52	0.03	0.50	0.54
		ShellWeight	0.53	0.03	0.51	0.55
		Diameter	0.5	0.03	0.48	0.52
		VisceraWeight	0.49	0.03	0.47	0.51
		ShuckledWeight	0.45	0.03	0.43	0.47
		WholeWeight	0.5	0.03	0.48	0.52
1000	Kendall's tau: t-test (7) probes do not eliminate any features	ShellWeight	0.52	0.02	0.51	0.53
		Height	0.51	0.02	0.50	0.52
		Diameter	0.5	0.02	0.49	0.51
		WholeWeight	0.49	0.02	0.48	0.50
		VisceraWeight	0.49	0.02	0.48	0.50
		ShuckledWeight	0.45	0.02	0.44	0.46
		Length	0.17	0.01	0.16	0.18
1000	Decision rule (3)	ShellWeight	0.53	0.03	0.51	0.55
		Length	0.17	0.01	0.16	0.18
		Gender	0.12	0.01	0.11	0.13

Table D.8: Abalone3C feature-feature correlations

Feature1	Feature2	corr_{ff}	Feature1	Feature2	corr_{ff}
Length	Diameter	0.92	Height	ShellWeight	0.79
Length	Height	0.75	WholeWeight	ShuckledWeight	0.88
Length	WholeWeight	0.88	WholeWeight	VisceraWeight	0.87
Length	ShuckledWeight	0.84	WholeWeight	ShellWeight	0.86
Length	VisceraWeight	0.83	ShuckledWeight	VisceraWeight	0.80
Length	ShellWeight	0.83	ShuckledWeight	ShellWeight	0.76
Diameter	Height	0.77	VisceraWeight	ShellWeight	0.80
Diameter	WholeWeight	0.88	Length	Gender	0.11
Diameter	ShuckledWeight	0.83	Diameter	Gender	0.46
Diameter	VisceraWeight	0.83	Height	Gender	0.47
Diameter	ShellWeight	0.85	WholeWeight	Gender	0.48
Height	WholeWeight	0.78	ShuckledWeight	Gender	0.46
Height	ShuckledWeight	0.72	VisceraWeight	Gender	0.49
Height	VisceraWeight	0.76	ShellWeight	Gender	0.47

Table D9: Feature selection for mushroom using SU coefficients

Sample size for SU measurement	Selection criteria (Number of selected features)	Selected features or top 5 features			
		Feature	Mean *SU*	StDev	95% CI of mean
500	t-test (4)	Ordor	0.55	0.03	0.02
		SporePrintColor	0.3	0.02	0.01
		RingType	0.23	0.01	0.01
		GillColor	0.2	0.02	0.01
500	Uniform probe (15) Uniform binary probe (14) Gaussian probe (21)	Ordor	0.55	0.03	0.02
		SporePrintColor	0.3	0.02	0.01
		StalkSfAbvRing	0.28	0.03	0.02
		GillSize	0.24	0.03	0.02
		StalkSfBlRing	0.23	0.03	0.02
500	Decision rule (14)	Ordor	0.55	0.03	0.02
		SporePrintColor	0.30	0.02	0.02
		StalkSfAbvRing	0.28	0.03	0.02
		GillSize	0.24	0.03	0.02
		StalkSfBlRing	0.23	0.03	0.02
		RingType	0.23	0.01	0.01
		GillColor	0.20	0.02	0.01
		StalkClAbvRing	0.18	0.02	0.01
		Bruises	0.17	0.03	0.02
		StalkClBlRing	0.15	0.02	0.01
		Population	0.14	0.02	0.01
		GillSpace	0.14	0.03	0.02
		habitat	0.11	0.01	0.01
		StalkRoot	0.10	0.01	0.01

Appendix E

Algorithm for breadth first generation of a search space

This appendix provides the details of the standard breadth-first search algorithm and the *BreadthFirstGenerate* algorithm which is based on the breadth first algorithm. The *BreadthFirstGenerate* algorithm was used for the generation of all possible tied predictions as discussed in section 6.4. The standard breadth-first search algorithm (Luger & Stubblefield, 1993) is given in figure E.1. The *BreadthFirstGenerate* algorithm is given in figure E.2.

Both algorithms use the lists OPEN, CLOSED and CHILDREN. The OPEN list holds the states that are still to be expanded. The CLOSED list holds all states that have been generated so far. The CHILDREN list is used to temporarily hold all the children (successor states) of the current state while the children are being generated. The major difference between the breadth-first-search algorithm and the BreadthFirstGenerate algorithm is that the breadth-first-search algorithm specifically searches for a goal state while the BreadthFirstGenerate algorithm simply generates all the possible states in the search space.

Breadth-first-search

```
1. OPEN = [start_state]
2. CLOSED = [ ]
3. while OPEN ≠ [ ]
   begin
          3.1  Remove leftmost state from OPEN, and call it X
          3.2  if X is the goal state
                    return X
          else
          3.2  generate children of X and put them on the CHILDREN list
          3.3 eliminate children of X on OPEN (prevent cycles)
          3.4  put X on CLOSED
          3.5 put all states on CHILDREN list on right end of OPEN
  end
```

Figure E.1: Breadth-first search algorithm

```
BreadthFirstGenerate( )
1. OPEN = [start_state]
2. CLOSED = [ ]
3. while OPEN ≠ [ ]
   begin
          3.1  Remove leftmost state from OPEN and  call it X
          3.2  generate children of X and put them on the CHILDREN list
          3.3  put X on CLOSED
          3.4 put all states on CHILDREN list on right end of OPEN
  end
```

Figure E.2: BreadthFirstGenerate algorithm

For the generation of all possible tied predictions, the predictions are assigned numbers 1,2,..,*k* corresponding to the *k* classes for the prediction task. The start state contains the first number (1). Each state {1,.., *j*} has the children *j*+1, *j*+2,..,,*k*. When the *BreadthFirstGenerate* algorithm has finished executing, all the possible states (tied predictions) are available on the CLOSED list.

Given a search space represented by a search tree with a constant branching factor *B*, the number of states (paths) of length *L* generated by a search algorithm is given by (Luger & Stubblefield, 1993: pg 146)

$$States = B + B^2 + B^3 + ... + B^L \qquad (E.1)$$

which reduces to:

$$States = B(B^L - 1)/(B - 1) \qquad (E.2)$$

For the *BreadthFirstGenerate* algorithm, the branching factor for level 1 of the tree is *k* -1 and reduces by 1 for successive levels. The maximum path length is *k* so that

$$States = (k-1) + (k-2)^2 + ... + (k-(k-1))^k \qquad (E.3)$$

which reduces to:

$$States = \sum_{j=1}^{k} (k-j)^j \qquad (E.4)$$

Appendix F

Predictive performance of single OVA and pVn models

The detailed results for predictive accuracy and TPRATE values for the single *k*-class, OVA aggregate and pVn aggregate models using the 5NN and See5 algorithms are provided in this appendix. Each table shows the accuracy and class TPRATE values for 10 test samples, as well as the mean, 95% confidence interval of the mean, standard deviation and variance. The mean values for performance were discussed in chapters 7 and 8. The variance values were used for the F-tests discussed in chapter 8.

F.1 5NN single 7-class and aggregate models for forest cover type

Tables F.1 to F.4 give the details of predictive accuracy and TRATE values for the 5NN single 7-class, OVA and pVn aggregate models forest cover type.

Table F.1: Predictive performance of the 5NN single 7- class model for forest cover type

Test set	Accuracy on all classes	5NN single model (equal class distribution) TPRATE% for class:						
		1	2	3	4	5	6	7
1	75.4	68	48	58	98	88	72	96
2	71.4	60	46	50	90	92	70	92
3	75.1	60	48	64	96	88	76	94
4	73.7	66	50	48	90	92	74	96
5	72.6	54	42	56	92	94	76	94
6	76.9	72	50	48	94	94	82	98
7	74.6	60	50	58	90	90	76	98
8	76	60	58	68	90	86	72	98
9	75.4	60	52	58	94	92	74	98
10	76	68	44	60	90	96	78	96
Mean	**74.7**	**62.8**	**48.8**	**56.8**	**92.4**	**91.2**	**75.0**	**96.0**
StDev	**1.7**	**5.4**	**4.4**	**6.6**	**3.0**	**3.2**	**3.4**	**2.1**
Variance	**2.9**	**29.5**	**19.7**	**43.7**	**8.7**	**10.0**	**11.8**	**4.4**
Mean & CI	**74.7±1.0**	**62.8±3.4**	**48.8±2.8**	**56.8±4.1**	**92.4±1.8**	**91.2±2.0**	**75.0±2.1**	**96.0±1.3**

Table F.2: Predictive performance of the 5NN un-boosted OVA aggregate model for forest cover type

Test set	Accuracy on all classes	5NN un-boosted OVA aggregate model. TPRATE% for class:						
		1	2	3	4	5	6	7
1	78.3	74	58	72	96	86	64	98
2	80.3	68	64	68	92	94	78	98
3	82.9	78	58	76	92	96	84	96
4	80.6	84	58	70	86	88	82	96
5	80.9	70	58	68	88	100	86	96
6	79.1	62	60	70	88	100	78	96
7	79.1	62	52	70	92	98	82	98
8	82.0	66	66	76	88	100	82	96
9	82.0	68	58	74	88	98	92	96
10	79.4	68	52	74	88	98	80	96
Mean	**80.5**	**70.0**	**58.4**	**71.8**	**89.8**	**95.8**	**80.8**	**96.6**
StDev	**1.5**	**6.9**	**4.4**	**3.0**	**3.0**	**5.0**	**7.2**	**1.0**
Variance	**2.3**	**48.0**	**19.4**	**9.3**	**9.3**	**25.3**	**51.7**	**0.9**
Mean&CI	**80.5±0.9**	**70±4.3**	**58.4±2.7**	**71.8±1.9**	**89.8±1.9**	**95.8±3.1**	**80.8±4.5**	**96.6±0.6**

TableF.3: Predictive performance of the 5NN boosted OVA aggregate model for forest cover type

Test set	Accuracy on all classes	5NN boosted OVA aggregate model. TPRATE% for class:						
		1	2	3	4	5	6	7
1	82.9	74	62	74	100	98	74	98
2	82.3	68	70	70	100	98	72	98
3	82.6	78	58	72	100	94	80	96
4	83.7	84	62	68	100	98	78	96
5	81.4	70	60	68	100	96	80	96
6	81.4	62	60	72	100	98	82	96
7	80.9	62	58	74	100	96	78	98
8	82.3	66	72	72	100	96	74	96
9	82.3	68	64	70	100	98	80	96
10	80.6	68	54	70	100	98	78	96
Mean	**82.0**	**70.0**	**62.0**	**71.0**	**100.0**	**97.0**	**77.6**	**96.6**
StDev	**1.0**	**6.9**	**5.5**	**2.2**	**0.0**	**1.4**	**3.2**	**1.0**
Variance	**0.9**	**48.0**	**30.2**	**4.7**	**0.0**	**2.0**	**10.5**	**0.9**
Mean &CI	**82.0±0.6**	**70.0±4.3**	**62.0±3.4**	**71.0±1.3**	**100.0±0.0**	**97.0±0.9**	**77.6±2.0**	**96.6±0.6**

Table F.4: Predictive performance of the 5NN pVn aggregate model for forest cover type

Test set	Accuracy on all classes	5NN pVn aggregate model. TPRATE% for class:						
		1	2	3	4	5	6	7
1	78.3	68	52	70	98	90	70	100
2	75.1	60	60	66	94	90	68	88
3	81.4	82	58	68	100	94	74	94
4	79.7	80	56	64	96	94	76	92
5	79.1	70	60	62	98	98	78	88
6	80.0	70	58	60	98	98	82	94
7	76.0	66	52	64	94	94	72	90
8	77.7	62	60	66	98	90	74	94
9	80.3	64	62	70	98	96	74	98
10	78.0	56	60	60	96	98	82	94
Mean	**78.6**	**67.8**	**57.8**	**65.0**	**97.0**	**94.2**	**75.0**	**93.2**
StDev	**2.0**	**8.2**	**3.5**	**3.7**	**1.9**	**3.3**	**4.6**	**3.9**
Variance	**3.8**	**68.0**	**12.0**	**13.6**	**3.8**	**11.1**	**21.6**	**15.3**
Mean&CI	**78.6±1.2**	**67.8±5.1**	**57.8±2.1**	**65.0±2.3**	**97.0±1.2**	**94.2±2.1**	**75.0±2.9**	**93.2±2.4**

F.2 See5 single 7-class and aggregate models for forest cover type

Tables F.5 to F.8 give the details of predictive accuracy and TPRATE values for the See5 single 7-class, OVA and pVn aggregate models for the forest cover type dataset.

Table F.5: Predictive performance of the See5 single 7-class model for forest cover type

Test set	Accuracy on **all classes**	See5 single model (equal class distribution). TPRATE% for class:						
		1	**2**	**3**	**4**	**5**	**6**	**7**
1	77.1	56	60	68	100	92	70	94
2	76	68	58	60	96	86	70	94
3	78	52	66	68	98	86	80	96
4	76	52	62	64	96	86	80	92
5	77.1	66	62	54	94	92	80	92
6	78.9	58	74	58	98	90	78	96
7	73.7	54	58	54	96	82	74	98
8	76	56	66	56	96	82	78	98
9	78.9	58	66	64	98	82	88	96
10	77.4	54	66	62	96	84	80	100
Mean	**76.91**	**57.40**	**63.80**	**60.80**	**96.80**	**86.20**	**77.80**	**95.60**
StDev	**1.57**	**5.50**	**4.85**	**5.27**	**1.69**	**3.94**	**5.37**	**2.63**
Variance	**2.47**	**30.27**	**23.51**	**27.73**	**2.84**	**15.51**	**28.84**	**6.93**
Mean & CI	**76.9±1.0**	**57.4±3.4**	**63.8±3.0**	**60.8±3.3**	**96.8±1.0**	**86.2±2.4**	**77.8±3.3**	**95.6±1.6**

Table F.6: Predictive performance of See5 un-boosted OVA aggregate model for forest cover type

Test sample	Accuracy on all classes	See5 un-boosted OVA aggregate model. TPRATE% for class:						
		1	2	3	4	5	6	7
1	74.9	64	44	68	84	96	80	88
2	75.7	62	52	66	88	94	78	90
3	75.1	60	50	60	92	90	78	96
4	73.7	60	40	62	88	96	80	90
5	74.3	60	44	64	86	98	78	90
6	77.1	68	58	60	86	98	80	90
7	74.6	58	50	66	88	94	72	94
8	75.1	54	52	66	82	96	80	100
9	76.9	64	50	66	86	92	82	98
10	75.4	56	58	62	86	90	84	92
Mean	**75.3**	**60.6**	**49.8**	**64.0**	**86.6**	**94.4**	**79.2**	**92.8**
StDev	**1.1**	**4.1**	**5.8**	**2.8**	**2.7**	**3.0**	**3.2**	**4.0**
Variance	**1.1**	**16.9**	**34.2**	**8.0**	**7.2**	**8.7**	**10.0**	**16.2**
Mean&CI	**75.3±0.7**	**60.6±2.6**	**49.8±3.6**	**64.0±1.8**	**86.6±1.7**	**94.4±1.8**	**79.2±2.0**	**92.8±2.5**

Table F.7: Predictive performance of See5 boosted OVA aggregate model for forest cover type

Test set	Accuracy on all classes	See5 boosted OVA aggregate model. TPRATE% for class:						
		1	**2**	**3**	**4**	**5**	**6**	**7**
1	79.4	70	70	72	96	82	70	96
2	80.3	68	66	70	94	92	74	98
3	80	60	66	66	100	90	78	100
4	77.7	62	66	62	94	86	78	96
5	78.9	60	70	60	96	92	78	96
6	80.3	70	76	54	96	90	78	98
7	78.9	62	74	60	96	90	72	98
8	78.6	66	70	60	92	90	76	96
9	80.6	72	66	62	94	90	80	100
10	79.1	60	74	66	96	82	76	100
Mean	**79.38**	**65.00**	**69.80**	**63.20**	**95.40**	**88.40**	**76.00**	**97.80**
StDev	**0.92**	**4.74**	**3.82**	**5.35**	**2.12**	**3.75**	**3.13**	**1.75**
Variance	**0.84**	**22.44**	**14.62**	**28.62**	**4.49**	**14.04**	**9.78**	**3.07**
Mean & CI	**79.4±0.6**	**65.0±2.9**	**69.8±2.4**	**63.2±3.3**	**95.4±1.3**	**88.4±2.3**	**76.0±1.9**	**97.8±1.1**

Table F.8: Predictive performance of the See5 pVn aggregate model for forest cover type

Test set	Accuracy on all classes	See5 pVn aggregate model. TPRATE% for class:						
		1	**2**	**3**	**4**	**5**	**6**	**7**
1	78	72	56	72	94	84	78	90
2	79.1	70	58	74	92	92	82	86
3	80.6	64	62	76	100	88	78	96
4	79.4	62	68	76	94	88	82	86
5	80	62	66	74	96	88	86	88
6	79.7	64	74	58	92	92	88	90
7	78.6	66	58	70	98	86	76	96
8	80.3	62	72	70	94	90	80	94
9	83.7	68	74	76	92	92	88	96
10	79.1	56	64	72	94	86	84	98
Mean	**79.85**	**64.60**	**65.20**	**71.80**	**94.60**	**88.60**	**82.20**	**92.00**
StDev	**1.56**	**4.62**	**6.75**	**5.37**	**2.67**	**2.84**	**4.26**	**4.52**
Variance	**2.44**	**21.38**	**45.51**	**28.84**	**7.16**	**8.04**	**18.18**	**20.44**
Mean& CI	**79.9±1.0**	**64.6±2.9**	**65.2±4.2**	**71.8±3.3**	**94.6±1.7**	**88.6±1.8**	**82.2±2.6**	**92.0±2.8**

F.3 5NN single 5-class and aggregate models for KDD Cup 1999

Tables F.9 to F.12 give the details of predictive accuracy and TPRATE values for the 5NN single 5-class, OVA and pVn aggregate models KDD Cup 1999.

Table F.9: Predictive performance of the 5NN single 5-class model for KDD Cup 1999

Test set	Accuracy on all classes	5NN single model (equal class distribution for NORMAL, DOS, PROBE, R2L). TPRATE% for class:				
		NORMAL	**DOS**	**PROBE**	**R2L**	**U2R**
1	69.7	81.4	80	95.7	60	31.4
2	72	87.1	72.9	97.1	71.4	31.4
3	65.7	87.1	51.4	98.6	60	31.4
4	71.1	94.3	61.4	94.3	72.9	32.9
5	68.3	81.4	62.9	92.9	72.9	31.4
6	66.3	85.7	60	94.3	60	31.4
7	69.7	87.1	71.4	94.3	64.3	31.4
8	66.3	81.4	67.1	94.3	57.1	31.4
9	69.7	82.8	71.4	98.6	64.3	31.4
10	66.6	75.7	64.3	97.1	64.3	31.4
Mean	**68.54**	**84.40**	**66.28**	**95.72**	**64.72**	**31.55**
StDev	**2.22**	**5.02**	**8.06**	**2.01**	**5.81**	**0.47**
Variance	**4.94**	**25.20**	**65.02**	**4.05**	**33.76**	**0.22**
Mean & CI	**68.5 ± 1.4**	**84.4 ± 3.1**	**66.3 ± 5.0**	**95.7 ± 1.2**	**64.7 ± 3.6**	**31.6 ± 0.3**

Table F.10: Predictive performance of the 5NN OVA un-boosted aggregate model for KDD Cup 1999

Test set	Accuracy on all classes	5NN un-boosted OVA aggregate model. TPRATE % for class:				
		NORMAL	DOS	PROBE	R2L	U2R
1	73.7	90	81.4	94.3	61.4	41.4
2	73.4	92.9	68.6	95.7	67.1	42.9
3	72.3	98.6	58.6	98.6	62.9	42.9
4	73.1	97.1	61.4	94.3	71.4	41.4
5	71.7	85.7	64.3	92.9	72.9	42.9
6	69.4	91.4	57.1	94.3	61.4	42.9
7	73.7	94.3	68.6	95.7	67.1	42.9
8	69.1	87.1	65.7	94.3	55.7	42.9
9	74.3	98.6	71.4	97.1	61.4	42.9
10	72.9	91.4	62.9	94.3	72.9	42.9
Mean	**72.4**	**92.7**	**66.0**	**95.2**	**65.4**	**42.6**
StDev	**1.8**	**4.5**	**7.1**	**1.7**	**5.8**	**0.6**
Variance	**3.2**	**20.3**	**49.7**	**2.8**	**33.6**	**0.4**
CI of mean	**1.1**	**2.8**	**4.4**	**1.0**	**3.6**	**0.4**
Mean&CI	**72.4±1.1**	**92.7±2.8**	**66.0±4.4**	**95.2±1.0**	**65.4±3.6**	**42.6±0.4**

Table F.11: Predictive performance of the 5NN OVA boosted aggregate model for KDD Cup 1999

Test set	Accuracy on all classes	5NN boosted OVA aggregate model. TPRATE% for class:				
		NORMAL	DOS	PROBE	R2L	U2R
1	73.7	90	82.9	94.3	58.6	42.9
2	73.4	94.3	68.6	95.7	65.7	42.9
3	70.0	98.6	52.9	98.6	57.1	42.9
4	72.3	97.1	61.4	94.3	65.7	42.9
5	70.9	85.7	64.3	92.9	74.3	37.1
6	68.0	90	58.6	94.3	57.1	40
7	71.4	94.3	70	95.7	58.6	38.6
8	68.3	85.7	67.1	94.3	55.7	38.6
9	72.3	98.6	71.4	98.6	54.3	38.6
10	70.0	90	62.9	95.7	61.4	40
Mean	**71.0**	**92.4**	**66.0**	**95.4**	**60.9**	**40.5**
StDev	**2.0**	**4.9**	**8.2**	**1.9**	**6.1**	**2.3**
Variance	**3.9**	**23.7**	**66.5**	**3.5**	**37.3**	**5.1**
CI of mean	**1.2**	**3.0**	**5.1**	**1.2**	**3.8**	**1.4**
Mean&CI	**71.0±1.2**	**92.4±3.0**	**66.0±5.1**	**95.4±1.2**	**60.9±3.8**	**40.5±1.4**

Table F.12: Predictive performance of the 5NN pVn aggregate model for KDD Cup 1999

Test sample	Accuracy on all classes	5NN pVn aggregate model. TPRATE% for class:				
		NORMAL	DOS	PROBE	R2L	U2R
1	79.4	97.1	100	98.6	72.9	28.6
2	82.0	98.6	98.6	98.6	88.6	27.1
3	80.6	100	95.7	100	81.4	25.7
4	82.0	100	98.6	98.6	85.7	25.7
5	81.0	100	98.6	97.1	85.7	25.7
6	78.0	95.7	94.3	100	77.1	21.4
7	83.0	100	98.6	98.6	87.1	31.4
8	80.0	98.6	100	97.1	74.3	28.6
9	77.1	98.6	91.4	100	72.9	22.9
10	80.0	98.6	97.1	95.7	88.6	20
Mean	**80.3**	**98.7**	**97.3**	**98.4**	**81.4**	**25.7**
StDev	**1.8**	**1.4**	**2.7**	**1.4**	**6.6**	**3.5**
Variance	**3.4**	**2.0**	**7.5**	**2.1**	**42.9**	**12.2**
Mean&CI	**80.3±1.1**	**98.7±0.9**	**97.3±1.7**	**98.4±0.9**	**81.4±4.1**	**25.7±2.2**

F.4 See5 single 5-class and aggregate models for KDD Cup 1999

Tables F.13 to F.16 give the details of predictive accuracy and TPRATE values for the See5 single 5-class, OVA and pVn aggregate models KDD Cup 1999.

Table F.13: Predictive performance of the See5 single model for KDD Cup 1999

Test set	Accuracy on all classes	See5 single model (equal class distribution for NORMAL, DOS, PROBE, R2L). TPRATE% for class:				
		NORMAL	**DOS**	**PROBE**	**R2L**	**U2R**
1	65.1	84.3	84.3	44.3	35.7	77.1
2	66.0	91.4	75.7	38.6	47.1	77.1
3	63.1	91.4	77.1	34.3	35.7	77.1
4	63.7	88.6	85.7	35.7	31.4	77.1
5	67.1	82.9	95.7	34.3	45.7	77.1
6	63.4	90.0	85.7	31.4	32.9	77.1
7	65.4	90.0	81.4	38.6	40.0	77.1
8	60.0	80.0	78.6	31.4	32.9	77.1
9	63.4	84.3	77.1	38.6	40.0	77.1
10	61.1	77.1	78.6	37.1	35.7	77.1
Mean	**63.83**	**86.00**	**81.99**	**36.43**	**37.71**	**77.10**
StDev	**2.17**	**5.03**	**6.07**	**3.90**	**5.38**	**0.00**
Variance	**4.72**	**25.30**	**36.84**	**15.19**	**28.97**	**0.00**
Mean & CI	**63.8±1.3**	**86.0±3.1**	**82.0±3.8**	**36.4±2.4**	**37.7±3.3**	**77.1±0.0**

Table F.14: Predictive performance of the See5 un-boosted OVA aggregate model for KDD Cup1999

Test set	Accuracy on all classes	See5 Class TPRATE% - boosted AGGREGATE MODEL				
		NORMAL	**DOS**	**PROBE**	**R2L**	**U2R**
1	62.3	97.1	45.7	88.6	34.3	45.7
2	65.7	100	54.3	87.1	41.4	45.7
3	60.9	98.6	42.9	85.7	31.4	45.7
4	66.6	98.6	61.4	88.6	38.6	45.7
5	64.9	98.6	54.3	84.3	41.1	45.7
6	61.1	97.1	37.1	91.4	34.3	45.7
7	64.3	98.6	55.7	87.1	34.3	45.7
8	62.3	97.1	51.4	90	27.1	45.7
9	62.6	100	52.9	88.6	25.7	45.7
10	62.3	97.1	45.7	88.6	34.3	45.7
Mean	**63.3**	**98.3**	**50.1**	**88.0**	**34.3**	**45.7**
StDev	**2.0**	**1.1**	**7.2**	**2.0**	**5.3**	**0.0**
Variance	**3.8**	**1.3**	**51.5**	**4.2**	**27.7**	**0.0**
Mean & CI	**63.3±1.2**	**98.3±0.7**	**50.1±4.4**	**88.0±1.3**	**34.3±3.3**	**45.7±0.0**

Table F.15: Predictive performance of the See5 boosted OVA aggregate model for KDD Cup1999

Test set	Accuracy on all classes	See5 boosted OVA aggregate model. TPRATE% for class:				
		NORMAL	**DOS**	**PROBE**	**R2L**	**U2R**
1	63.1	97.1	65.7	88.6	24.3	40.0
2	63.7	100.0	61.4	90.0	27.1	40.0
3	60.9	100.0	50.0	88.6	25.7	40.0
4	61.7	100.0	61.4	90.0	17.1	40.0
5	61.4	98.6	54.3	84.3	30.0	40.0
6	59.1	98.6	42.9	92.9	21.4	40.0
7	62.9	100.0	61.4	88.6	24.3	40.0
8	60.3	98.6	52.9	90.0	20.0	40.0
9	61.1	100.0	60.0	91.4	14.3	40.0
10	62.3	98.6	52.9	88.6	31.4	40.0
Mean	**61.65**	**99.15**	**56.29**	**89.30**	**23.56**	**40.00**
StDev	**1.40**	**1.00**	**6.88**	**2.26**	**5.44**	**0.00**
Variance	**1.95**	**1.00**	**47.38**	**5.09**	**29.55**	**0.00**
Mean & CI	**61.7±0.9**	**99.2±0.6**	**56.3±4.3**	**89.3±1.4**	**23.6±3.4**	**40.0±0.0**

Table F.16: Predictive performance of the See5 pVn aggregate model for KDD Cup 1999

Test sample ID	Accuracy on all classes	See5 pVn aggregate model. TPRATE% for class:				
		NORMAL	**DOS**	**PROBE**	**R2L**	**U2R**
1	74	97.1	67.1	98.6	30	77.1
2	79.1	98.6	57.1	97.1	65.7	77.1
3	78	98.6	60	97.1	57.1	77.1
4	83.4	98.6	87.1	95.7	58.6	77.1
5	85.1	100	84.3	97.1	67.1	77.1
6	78	97.1	64.3	100	51.4	77.1
7	81.1	98.6	71.4	95.7	62.9	77.1
8	77.7	97.1	70	97.1	47.1	77.1
9	74.9	98.6	55.7	97.1	45.7	77.1
10	78.3	97.1	67.1	94.3	55.7	77.1
Mean	**78.96**	**98.14**	**68.41**	**96.98**	**54.13**	**77.10**
StDev	**3.45**	**0.99**	**10.51**	**1.57**	**11.16**	**0.00**
Variance	**11.88**	**0.98**	**110.42**	**2.48**	**124.50**	**0.00**
Mean & CI	**79.0 ± 2.1**	**98.1 ± 0.6**	**68.4 ± 6.5**	**97.0 ± 1.0**	**54.1 ± 6.9**	**77.1 ± 0.0**

F.5 Single and aggregate models for wine quality (white)

Tables F.17 through F.24 give the details of predictive accuracy and TPRATE values for the 5NN single and aggregate models for the wine quality (white) dataset. Tables F.25 and F.26 provide the statistical test results for the comparison of the single and aggregate models.

Table F.17: Predictive performance of the 5NN single model for Wine quality

Test set	Accuracy on all classes	5NN single model TPRATE% for class:				
		4	5	6	7	8
1	31.2	8	56	22	54	8
2	30	14	58	30	44	4
3	29.2	10	56	24	44	12
4	28.8	6	54	30	46	8
5	33.2	14	54	34	54	10
6	32.4	12	54	34	54	8
7	30.8	14	46	36	44	14
8	34	18	50	36	54	12
9	35.2	10	64	38	50	14
10	31.6	10	56	30	50	12
Mean	**31.6**	**11.6**	**54.8**	**31.4**	**49.4**	**10.2**
StDev	**2.1**	**3.5**	**4.7**	**5.3**	**4.5**	**3.2**
Variance	**4.3**	**12.3**	**22.4**	**27.6**	**20.5**	**10.2**
Mean & CI	**31.6±1.3**	**11.6±2.2**	**54.8±2.9**	**31.4±3.3**	**49.4±2.8**	**10.2±2.0**

Table F.18: Predictive performance of the 5NN un-boosted OVA model for Wine quality

Test set	Accuracy on all classes	5NN OVA un-boosted model TPRATE% for class:				
		4	5	6	7	8
1	30.4	16	54	24	50	8
2	32.8	14	60	28	56	6
3	35.2	10	64	34	54	14
4	28.8	6	58	32	40	8
5	33.6	14	60	34	52	8
6	30.8	14	54	32	48	6
7	30.4	12	58	26	44	12
8	34	18	58	30	52	12
9	32.8	10	56	30	54	14
10	32.8	12	68	24	48	12
Mean	**32.2**	**12.6**	**59.0**	**29.4**	**49.8**	**10.0**
StDev	**1.9**	**3.4**	**4.3**	**3.8**	**4.9**	**3.1**
Variance	**3.8**	**11.6**	**18.9**	**14.3**	**24.4**	**9.8**
Mean & CI	**32.2±1.2**	**12.6±2.1**	**59.0±2.7**	**29.9±2.3**	**49.8±3.1**	**10.0±1.9**

Table F.19: Predictive performance of the 5NN boosted OVA model for Wine quality

Test set	Accuracy on all classes	5NN OVA boosted model TPRATE% for class:				
		4	5	6	7	8
1	33.2	16	64	24	52	10
2	33.2	16	68	16	58	8
3	35.2	12	68	26	54	16
4	29.2	6	66	22	44	8
5	35.2	16	64	34	52	10
6	29.6	14	62	16	50	6
7	35.2	14	68	34	44	16
8	35.6	18	62	28	56	14
9	34.4	12	64	22	60	14
10	34.8	12	70	28	52	12
Mean	**33.6**	**13.6**	**65.6**	**25.0**	**52.2**	**11.4**
StDev	**2.3**	**3.4**	**2.8**	**6.3**	**5.3**	**3.5**
Variance	**5.5**	**11.4**	**7.8**	**40.2**	**28.0**	**12.5**
Mean & CI	**33.6±1.5**	**13.6±2.1**	**65.6±1.7**	**25.0±3.9**	**52.2±3.3**	**11.4±2.2**

Table F.20: Predictive performance of the 5NN pVn model for Wine quality

Test set	Accuracy on all classes	5NN pVn aggregate model TPRATE% for class:				
		4	5	6	7	8
1	33.2	16	44	52	46	8
2	34.8	12	56	50	50	6
3	36	12	58	50	46	14
4	31.2	4	54	58	32	8
5	37.6	14	60	60	44	10
6	32.4	12	54	54	34	8
7	32.4	10	56	46	36	14
8	35.6	18	52	42	54	12
9	34	6	50	50	50	14
10	38.4	10	66	54	50	12
Mean	**34.6**	**11.4**	**55.0**	**51.6**	**44.2**	**10.6**
StDev	**2.4**	**4.2**	**5.9**	**5.3**	**7.6**	**3.0**
Variance	**5.6**	**17.8**	**34.9**	**28.3**	**58.2**	**8.9**
Mean & CI	**34.6±1.5**	**11.4±2.6**	**55.0±3.7**	**51.6±3.3**	**44.2±4.7**	**10.6±1.9**

Table F.21: Predictive performance of the See5 single model for Wine quality

Test set	Accuracy on all classes	See5 single model TPRATE% for class:				
		4	5	6	7	8
1	38.4	28	70	32	54	8
2	37.6	24	70	34	52	8
3	38.4	28	74	32	50	8
4	33.6	20	64	26	46	12
5	36.4	28	70	32	48	4
6	37.2	30	72	30	46	8
7	36.8	28	70	36	44	6
8	37.2	28	66	36	46	10
9	38	26	70	34	50	10
10	34	20	74	30	42	
Mean	**36.8**	**26.0**	**70.0**	**32.2**	**47.8**	**8.2**
StDev	**1.7**	**3.5**	**3.1**	**3.0**	**3.7**	**2.3**
Variance	**2.9**	**12.4**	**9.8**	**9.3**	**13.7**	**5.4**
Mean & CI	**36.8±1.0**	**26.0±2.2**	**70.0±1.9**	**32.2±1.9**	**47.8±2.3**	**8.2±1.4**

Table F.22: Predictive performance of the See5 un-boosted model for Wine quality

Test set	Accuracy on all classes	See5 un-boosted OVA model TPRATE% for class:				
		4	5	6	7	8
1	34	42	64	14	36	14
2	34.8	38	68	20	40	8
3	38	42	70	12	48	18
4	29.6	26	68	6	42	6
5	36.4	48	70	10	42	12
6	32	34	66	10	40	10
7	36	46	58	14	46	16
8	38	46	60	18	52	14
9	34	40	58	14	38	20
10	35.6	40	64	16	44	14
Mean	**34.8**	**40.2**	**64.6**	**13.4**	**42.8**	**13.2**
StDev	**2.6**	**6.5**	**4.6**	**4.1**	**4.8**	**4.3**
Variance	**6.8**	**42.2**	**21.4**	**16.9**	**23.3**	**18.8**
Mean & CI	**34.8±1.6**	**40.2±4.0**	**64.6±2.9**	**13.4±2.6**	**42.8±3.0**	**13.2±2.7**

Table F.23: Predictive performance of the See5 boosted model for Wine quality

Test set	Accuracy on all classes	See5 boosted OVA model TPRATE% for class:				
		4	5	6	7	8
1	36.4	42	68	14	42	16
2	36	38	72	16	46	8
3	37.6	42	74	6	48	18
4	31.2	26	72	4	46	8
5	36.4	48	72	6	42	14
6	33.2	34	66	10	44	12
7	36.4	46	62	8	50	16
8	38.8	46	64	14	56	14
9	35.2	40	62	12	42	20
10	34.8	40	66	8	46	14
Mean	**35.6**	**40.2**	**67.8**	**9.8**	**46.2**	**14.0**
StDev	**2.2**	**6.5**	**4.5**	**4.0**	**4.4**	**3.9**
Variance	**4.7**	**42.2**	**20.0**	**16.4**	**19.1**	**15.1**
Mean & CI	**35.6±1.3**	**40.2±4.0**	**67.8±2.8**	**9.8±2.5**	**46.2±2.7**	**14.0±2.4**

Table F.24: Predictive performance of the See5 pVn model for Wine quality

Test set	Accuracy on all classes	See5 pVn model TPRATE% for class:				
		4	5	6	7	8
1	42	34	56	44	60	16
2	39.6	28	54	42	64	10
3	42.4	40	58	36	64	14
4	38	26	58	38	58	10
5	40.8	38	56	42	54	14
6	39.2	38	50	42	50	16
7	42.8	34	62	48	54	16
8	41.2	38	50	40	64	14
9	38.8	32	46	42	56	18
10	40.8	36	60	32	60	16
Mean	**40.6**	**34.4**	**55.0**	**40.6**	**58.4**	**14.4**
StDev	**1.6**	**4.6**	**5.0**	**4.4**	**4.9**	**2.6**
Variance	**2.6**	**21.2**	**25.1**	**19.6**	**23.8**	**6.9**
Mean & CI	**40.6±1.0**	**34.4±2.9**	**55.0±3.1**	**40.6±2.7**	**58.4±3.0**	**14.4±1.6**

Table F.25: Statistical tests for 5NN single and aggregate model comparison for wine quality

Wine quality white: 5NN models						
Group names and mean accuracy /TPRATE:10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A aggregate model	Group B single model	95% CI of mean difference	P value (2 tail)	Group A better than Group B?	Diff(A,B)%	Ratio(A,B)
OVA un-boosted All classes-A (32.2 ± 1.2)	All classes-S (31.6 ± 1.3)	[-1.2, 2.2]	0.511	no	0.5	0.01
OVA un-boosted Class4-A (12.6 ± 2.1)	Class4-S (11.6 ± 2.2)	[-0.9, 2.9]	0.273	no	1.0	0.01
OVA un-boosted Class5-A 59.0 ± 2.7)	Class5-S (54.8 ± 2.9)	[-0.3, 8.7]	0.066	yes	4.2	0.09
OVA un-boosted Class6-A (29.9 ± 2.3)	Class6-S (31.4 ± 3.3)	[-6.2, 2.2]	0.311	no	-2.0	-0.03
OVA un-boosted Class7-A (49.8 ± 3.1)	Class7-S (49.4 ± 2.8)	[-4.1, 4.9]	0.846	no	0.4	0.01
OVA un-boosted Class8-A (10.0 ± 1.9)	Class8-S (10.2 ± 2.0)	[-1.3, 0.9]	0.678	no	-0.2	0.00
OVA boosted All classes-A (33.6 ± 1.5)	All classes-S (31.6 ± 1.3)	[0.1, 3.7]	0.041	yes	1.9	0.03
OVA boosted Class4-A 13.6 ± 2.1)	Class4-S (11.6 ± 2.2)	[0.3, 3.7]	0.023	yes	2.0	0.02
OVA boosted Class5-A (65.6 ± 1.7)	Class5-S (54.8 ± 2.9)	[6.9, 14.7]	0.000	yes	10.8	0.24
OVA boosted Class6-A (25.0 ± 3.9)	Class6-S (31.4 ± 3.3)	[-11.8, -1.0]	0.025	no	-6.4	-0.09
OVA boosted Class7-A (52.2 ± 3.3)	Class7-S (49.4 ± 2.8)	[-1.6, 7.3]	0.191	no	2.8	0.06
OVA un-boosted Class8-A (11.4 ± 2.2)	Class8-S (10.2 ± 2.0)	[-0.2, 2.9]	0.081	yes	1.2	0.01
pVn All classes-A (34.6 ± 1.5)	All classes-S (31.6 ± 1.3)	[1.0, 4.9]	0.008	yes	2.9	0.04
pVn Class4-A (11.4 ± 2.6)	Class4-S (11.6 ± 2.2)	[-2.7, 2.3]	0.859	no	-0.2	0.00
pVn Class5-A (55.0 ± 3.7)	Class5-S (54.8 ± 2.9)	-5.6, 6.0]	0.939	no	0.2	0.00
pVn Class6-A (51.6 ± 3.3)	Class6-S (31.4 ± 3.3)	[14.3, 26.1]	0.000	yes	20.2	0.29
pVn Class7-A (44.2 ± 4.7)	Class7-S (49.4 ± 2.8)	[-11.0, 0.6]	0.074	no	-5.2	-0.10
pVn Class8-A (10.6 ± 1.9)	Class8-S (10.2 ± 2.0)	[-0.2, 1.0]	0.168	no	0.4	0.00

Table F.26: Statistical tests for See5 single and aggregate model comparison for wine quality

Wine quality white: See5 models						
Group names and mean accuracy /TPRATE:10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A aggregate model	Group B single model	95% CI of mean difference	P value (2 tail)	Group A better than Group B?	Diff(A,B)%	Ratio(A,B)
OVA un-boosted All classes-A (34.8 ± 1.6)	All classes-S (36.8 ± 1.0)	[-3.7, -0.2]	0.034	no	-1.9	-0.03
OVA un-boosted Class4-A (40.2 ± 4.0)	Class4-S (26.0 ± 2.2)	[10.3, 18.1]	0.000	yes	14.2	0.19
OVA un-boosted Class5-A 64.6 ± 2.9)	Class5-S (70.0 ± 1.9)	[-9.1, -1.7]	0.009	no	-5.4	-0.18
OVA un-boosted Class6-A (13.4 ± 2.9)	Class6-S (32.2 ± 1.9)	[-20.8, -16.8]	0.000	no	-18.8	-0.28
OVA un-boosted Class7-A (42.8 ± 3.8)	Class7-S (47.8 ± 2.8)	[-10.3, 0.3]	0.062	no	-5.0	-0.10
OVA un-boosted Class8-A (13.2 ± 2.7)	Class8-S (8.2 ± 1.4)	[0.7, 9.1]	0.028	yes	5.0	0.05
OVA boosted All classes-A (35.6 ± 1.3)	All classes-S (36.8 ± 1.0)	[-2.4, 0.1]	0.062	no	-1.2	-0.02
OVA boosted Class4-A (40.2 ± 4.0)	Class4-S (26.0 ± 2.2)	[10.3, 18.1]	0.000	yes	14.2	0.19
OVA boosted Class5-A (67.8 ± 2.8)	Class5-S (70.0 ± 1.9)	[-6.0, 1.6]	0.227	no	-2.2	-0.07
OVA boosted Class6-A (9.8 ± 2.5)	Class6-S (32.2 ± 1.9)	[-24.8, -20.0]	0.000	no	-22.4	-0.33
OVA boosted Class7-A (46.2 ± 2.7)	Class7-S (47.8 ± 2.8)	[-6.5, 3.3]	0.475	no	-1.6	-0.03
OVA un-boosted Class8-A (14.0 ± 2.4)	Class8-S (8.2 ± 1.4)	[1.8, 9.7]	0.100	yes	5.8	0.06
pVn All classes-A (40.6 ± 1.0)	All classes-S (36.8 ± 1.0)	[2.4, 5.1]	0.000	yes	3.8	0.06
pVn Class4-A (34.4 ± 2.9)	Class4-S (26.0 ± 2.2)	[5.8, 11.0]	0.000	yes	8.4	0.11
pVn Class5-A (55.0 ± 3.1)	Class5-S (70.0 ± 1.9)	[-18.9, -11.1]	0.000	no	-15.0	-0.50
pVn Class6-A (40.6 ± 2.7)	Class6-S (32.2 ± 1.9)	[5.6, 11.2]	0.000	yes	8.4	0.12
pVn Class7-A (58.4 ± 3.0)	Class7-S (47.8 ± 2.8)	[7.0, 14.2]	0.000	yes	10.6	0.20
pVn Class8-A (14.4 ± 1.6]	Class8-S (8.2 ± 1.4)	[2.9, 9.1]	0.002	yes	6.2	0.07

Appendix G

ROC analysis details

The computational method for the AUC and the detailed results for ROC analysis are provided in this appendix. The ROC analysis that was conducted for the experiments was discussed in chapter 9. The method used to compute the Area Under the ROC curve (AUC) is depicted in figure G.1 and table G.1. Figure G.1 shows a ROC curve created with three points corresponding to three threshold points λ_1, λ_2 and λ_3. The x-axis and y-axis respectively represent the FPRATE and TPRATE of a probabilistic classifier. Threshold averaging was used for the computation of the AUC. Recall from chapter 9 that for threshold averaging, the co-ordinates of each point on the ROC curve are obtained by computing the mean FPRATE (x co-ordinate) and mean TPRATE (y co-ordinate) for one threshold value λ_i. The mean FPRATE and TPRATE values were computed for 10 test sets. The areas of regions A1 to A7 were used to compute the AUC as shown in table G.1.

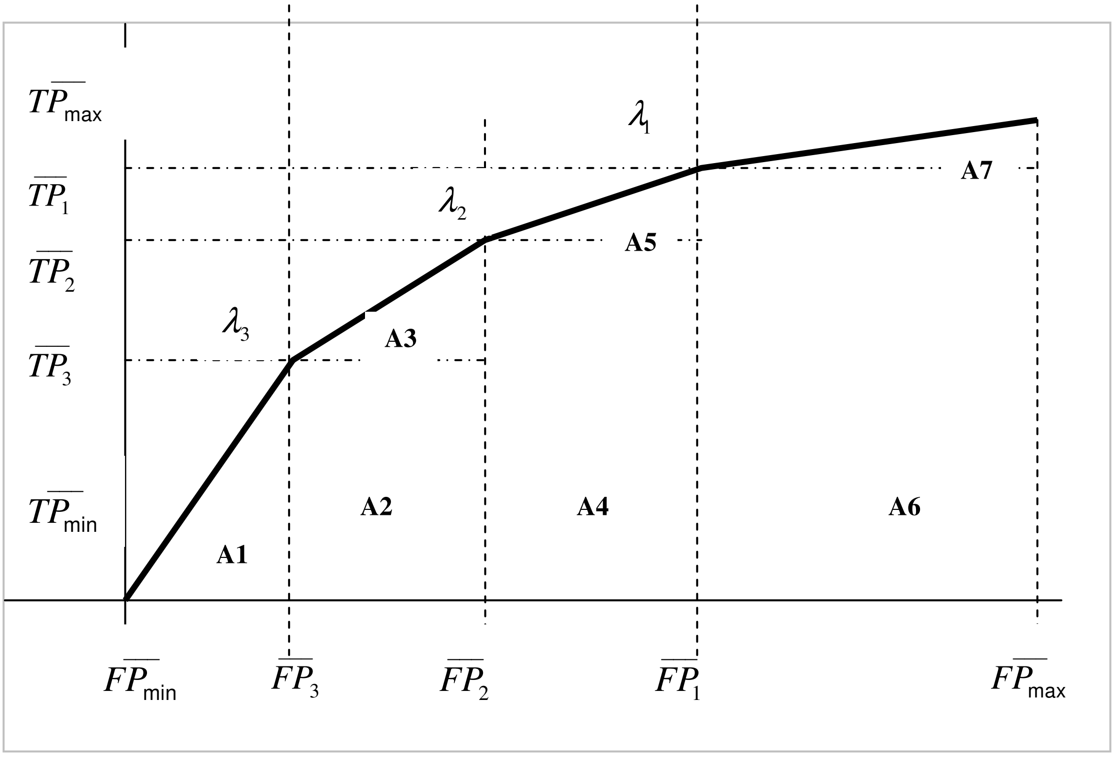


Figure G.1: Areas of the ROC plane used to compute the AUC

Table G.1: Method used for the computation of the AUC for probabilistic classifiers

Area code	Computation
A1	½ * (FP3 * TP3)
A2	(FP2 – FP3) * TP3
A3	½ * (FP2 – FP3) * (TP2 – TP3)
A4	(FP1 – FP2) * TP2
A5	½ * (FP1 – FP2) * (TP1 - TP2)
A6	(FPmax - FP1) * TP1
A7	½ * (FPmax - FP1) * (TPmax – TP1)
TOTAL	A1 + A2 + A3 + A4 + A5 + A6 + A7
AUC_{above}	(TOTAL – area under 45deg line)
AUC	TOTAL

Tables G.2 to G.7 provide the details of the FPRATE values (FP1, FP2, FP3) and TPRATE values (TP1, TP2, TP3) and AUC values for the forest cover type, KDD Cup 1999 and wine quality datasets. The AUC is the area between the x-axis, y-axis and ROC curve. AUC_{above} is the area between the 45 degree line and the ROC curve. The threshold values of 0.6, 0.8 and 1.0 for the 5NN classifiers correspond to the number of nearest neighbours (3, 4, 5) used by the 5NN algorithm to determine the winning class. The threshold values of 0.5, 0.75 and 1.0 were used for the See5 classifiers. The positive class column represents a *one-vs-rest* classifier which predicts the indicated class as the positive class and all the other classes as negative classes.

Table G.2: One-vs-rest AUC for the 5NN forest cover type models

5NN forest cover type models: TPRATE, FPRATE, AUC and Mean AUC									
		Mean values for thresholds							
		$\lambda_1 = 0.6$		$\lambda_2 = 0.8$		$\lambda_3 = 1.0$			
Model	Positive class	FP1	TP1	FP2	TP2	FP3	TP3	*AUC*	AUC_{above}
single 5NN	1	0.04	0.62	0.02	0.37	0.00	0.17	0.79	0.29
	2	0.03	0.48	0.01	0.27	0.00	0.09	0.73	0.23
	3	0.03	0.51	0.01	0.32	0.00	0.15	0.75	0.25
	4	0.03	0.92	0.02	0.78	0.01	0.53	0.95	0.45
	5	0.03	0.88	0.02	0.78	0.01	0.48	0.93	0.43
	6	0.05	0.70	0.02	0.44	0.01	0.18	0.83	0.33
	7	0.03	0.95	0.01	0.82	0.01	0.64	0.97	0.47
						Mean:		**0.85**	**0.35**
OVA unboosted 5NN	1	0.03	0.70	0.03	0.69	0.03	0.58	0.83	0.33
	2	0.03	0.58	0.03	0.57	0.02	0.49	0.78	0.28
	3	0.03	0.72	0.03	0.72	0.02	0.60	0.85	0.35
	4	0.02	0.90	0.02	0.87	0.01	0.67	0.94	0.44
	5	0.04	0.96	0.04	0.96	0.03	0.89	0.96	0.46
	6	0.05	0.81	0.05	0.80	0.03	0.67	0.88	0.38
	7	0.03	0.97	0.03	0.97	0.02	0.91	0.97	0.47
						Mean:		**0.89**	**0.39**
OVA boosted 5NN	1	0.03	0.70	0.03	0.69	0.03	0.58	0.83	0.33
	2	0.03	0.62	0.03	0.60	0.02	0.51	0.80	0.30
	3	0.03	0.71	0.03	0.71	0.02	0.61	0.84	0.34
	4	0.02	1.00	0.02	1.00	0.01	1.00	0.99	0.49
	5	0.04	0.97	0.03	0.94	0.02	0.82	0.97	0.47
	6	0.04	0.78	0.04	0.75	0.03	0.63	0.87	0.37
	7	0.03	0.97	0.03	0.97	0.02	0.91	0.97	0.47
						Mean:		**0.90**	**0.40**
pVn 5NN	1	0.05	0.68	0.03	0.62	0.01	0.36	0.82	0.32
	2	0.04	0.57	0.03	0.50	0.02	0.30	0.77	0.27
	3	0.04	0.65	0.03	0.52	0.01	0.34	0.81	0.31
	4	0.03	0.97	0.02	0.83	0.01	0.68	0.98	0.48
	5	0.04	0.94	0.03	0.89	0.02	0.79	0.96	0.46
	6	0.05	0.75	0.03	0.65	0.01	0.39	0.86	0.36
	7	0.02	0.93	0.01	0.67	0.01	0.67	0.96	0.46
						Mean:		**0.88**	0.38

Table G.3: One-vs-rest AUC for the 5NN KDD Cup 1999 models

5NN KDD Cup 1999 models: TPRATE, FPRATE, AUC and Mean AUC									
		Mean values for thresholds							
		$\lambda_1 = 0.6$		$\lambda_2 = 0.8$		$\lambda_3 = 1.0$			
Model	Positive class	FP1	TP1	FP2	TP2	FP3	TP3	AUC	AUC_{above}
single 5NN	NORM	0.22	0.84	0.13	0.84	0.11	0.80	0.86	0.36
	R2L	0.06	0.65	0.05	0.60	0.04	0.53	0.80	0.30
	DOS	0.01	0.66	0.01	0.63	0.01	0.61	0.83	0.33
	PROBE	0.09	0.96	0.09	0.96	0.07	0.93	0.94	0.44
	U2R	0.01	0.31	0.01	0.26	0.01	0.20	0.65	0.15
						Mean:		**0.82**	**0.32**
OVA unboosted 5NN	NORM	0.14	0.92	0.13	0.92	0.10	0.92	0.91	0.41
	R2L	0.07	0.65	0.07	0.62	0.06	0.57	0.79	0.29
	DOS	0.00	0.66	0.00	0.66	0.00	0.65	0.83	0.33
	PROBE	0.08	0.95	0.08	0.95	0.08	0.95	0.94	0.44
	U2R	0.01	0.43	0.01	0.43	0.01	0.31	0.71	0.21
						Mean:		**0.83**	**0.33**
OVA boosted 5NN	NORM	0.15	0.92	0.13	0.92	0.10	0.92	0.91	0.41
	R2L	0.07	0.61	0.06	0.59	0.05	0.52	0.77	0.27
	DOS	0.00	0.66	0.00	0.66	0.00	0.66	0.83	0.33
	PROBE	0.08	0.95	0.08	0.95	0.08	0.95	0.94	0.44
	U2R	0.01	0.40	0.01	0.40	0.01	0.29	0.70	0.20
						Mean:		**0.83**	**0.33**
pVn 5NN	NORM	0.16	0.99	0.15	0.99	0.12	0.98	0.93	0.43
	R2L	0.07	0.81	0.06	0.81	0.06	0.78	0.88	0.38
	DOS	0.00	0.97	0.00	0.97	0.00	0.72	0.98	0.48
	PROBE	0.00	0.98	0.00	0.98	0.00	0.98	0.99	0.49
	U2R	0.01	0.26	0.01	0.20	0.00	0.05	0.63	0.13
						Mean:		**0.88**	**0.33**

Table G.4: One-vs-rest AUC for the 5NN Wine quality models

5NN Wine quality (white) models: Mean TPRATE, mean FPRATE, AUC and Mean AUC									
		Mean values for thresholds							
		$\lambda_1 = 0.6$		$\lambda_2 = 0.8$		$\lambda_3 = 1.0$			
Model	positive class	FP1	TP1	FP2	TP2	FP3	TP3	AUC	AUC_{above}
single 5NN	4	0.04	0.12	0.02	0.08	0.01	0.04	0.54	0.04
	5	0.20	0.48	0.10	0.29	0.04	0.11	0.65	0.15
	6	0.17	0.22	0.04	0.08	0.01	0.01	0.53	0.03
	7	0.23	0.42	0.10	0.16	0.04	0.03	0.59	0.09
	8	0.02	0.10	0.01	0.07	0.00	0.03	0.54	0.04
						Mean	**AUC:**	**0.57**	**0.07**
OVA un-boosted 5NN	4	0.05	0.13	0.05	0.13	0.04	0.09	0.54	0.04
	5	0.27	0.59	0.25	0.55	0.12	0.34	0.67	0.17
	6	0.22	0.29	0.19	0.26	0.11	0.16	0.54	0.04
	7	0.28	0.50	0.25	0.42	0.17	0.33	0.61	0.11
	8	0.02	0.10	0.02	0.10	0.02	0.10	0.54	0.04
						Mean	**AUC:**	**0.58**	**0.08**
OVA boosted 5NN	4	0.05	0.14	0.05	0.14	0.05	0.10	0.54	0.04
	5	0.31	0.66	0.29	0.59	0.13	0.35	0.68	0.18
	6	0.13	0.25	0.09	0.20	0.02	0.08	0.56	0.06
	7	0.29	0.52	0.27	0.48	0.17	0.35	0.62	0.12
	8	0.03	0.11	0.03	0.11	0.02	0.11	0.54	0.04
						Mean	**AUC:**	**0.59**	**0.09**
pVn 5NN	4	0.05	0.11	0.04	0.09	0.02	0.02	0.53	0.03
	5	0.23	0.53	0.15	0.39	0.04	0.16	0.66	0.16
	6	0.27	0.50	0.17	0.32	0.04	0.12	0.62	0.12
	7	0.22	0.44	0.15	0.28	0.06	0.08	0.60	0.10
	8	0.02	0.11	0.02	0.09	0.01	0.06	0.55	0.05
						Mean	**AUC:**	**0.59**	**0.09**

Table G.5: One-vs-rest AUC for the See5 forest cover type models

See5 forest cover type models: TPRATE, FPRATE, AUC and Mean AUC									
Model	Positive class	Mean values for thresholds						AUC	AUC_{above}
		$\lambda_1 = 0.5$		$\lambda_2 = 0.75$		$\lambda_3 = 1.0$			
		FP1	TP1	FP2	TP2	FP3	TP3		
single See5	1	0.03	0.57	0.01	0.28	0.00	0.04	0.77	0.27
	2	0.06	0.63	0.03	0.39	0.00	0.04	0.79	0.29
	3	0.03	0.61	0.01	0.41	0.00	0.04	0.79	0.29
	4	0.03	0.94	0.02	0.90	0.00	0.08	0.96	0.46
	5	0.03	0.86	0.02	0.77	0.00	0.00	0.92	0.42
	6	0.05	0.78	0.03	0.60	0.00	0.05	0.87	0.37
	7	0.03	0.96	0.02	0.85	0.01	0.03	0.97	0.47
						Mean:		**0.87**	**0.37**
OVA unboosted 5NNSee5	1	0.05	0.61	0.05	0.60	0.00	0.01	0.78	0.28
	2	0.05	0.50	0.05	0.50	0.00	0.00	0.72	0.22
	3	0.04	0.64	0.04	0.62	0.00	0.02	0.80	0.30
	4	0.01	0.87	0.01	0.85	0.00	0.00	0.93	0.43
	5	0.04	0.94	0.04	0.94	0.00	0.01	0.95	0.45
	6	0.07	0.79	0.07	0.79	0.00	0.08	0.86	0.36
	7	0.03	0.93	0.03	0.93	0.00	0.00	0.95	0.45
						Mean:		**0.86**	**0.36**
OVA boosted See5	1	0.03	0.63	0.02	0.52	0.00	0.04	0.80	0.30
	2	0.07	0.67	0.07	0.62	0.01	0.01	0.80	0.30
	3	0.02	0.63	0.02	0.62	0.00	0.08	0.80	0.30
	4	0.01	0.95	0.01	0.94	0.00	0.04	0.97	0.47
	5	0.04	0.87	0.04	0.87	0.00	0.09	0.92	0.42
	6	0.04	0.76	0.04	0.76	0.00	0.05	0.86	0.36
	7	0.01	0.98	0.01	0.97	0.00	0.22	0.98	0.48
						Mean:		**0.88**	**0.38**
pVn See5	1	0.04	0.65	0.02	0.54	0.00	0.08	0.81	0.31
	2	0.06	0.65	0.05	0.61	0.00	0.04	0.80	0.30
	3	0.04	0.72	0.03	0.68	0.00	0.09	0.84	0.34
	4	0.01	0.95	0.01	0.89	0.00	0.01	0.97	0.47
	5	0.02	0.89	0.02	0.81	0.00	0.00	0.93	0.43
	6	0.05	0.82	0.04	0.78	0.00	0.12	0.89	0.39
	7	0.02	0.92	0.01	0.88	0.00	0.03	0.95	0.45
						Mean:		**0.88**	**0.38**

Table G.6: One-vs-rest AUC for the See5 KDD Cup 1999 models

See5 KDD Cup 1999 models: TPRATE, FPRATE, AUC and Mean AUC									
		Mean values for thresholds							
		$\lambda_1 = 0.5$		$\lambda_2 = 0.75$		$\lambda_3 = 1.0$			
Model	Positive class	FP1	TP1	FP2	TP2	FP3	TP3	*AUC*	AUC_{above}
single See5	NORM	0.22	0.86	0.22	0.86	0.02	0.63	0.88	0.38
	R2L	0.02	0.38	0.02	0.38	0.00	0.12	0.68	0.18
	DOS	0.02	0.82	0.02	0.82	0.02	0.82	0.90	0.40
	PROBE	0.04	0.36	0.04	0.36	0.02	0.36	0.67	0.17
	U2R	0.16	0.77	0.16	0.77	0.00	0.00	0.81	0.31
						Mean:		**0.79**	**0.29**
OVA unboosted See5	NORM	0.11	0.98	0.11	0.98	0.10	0.98	0.94	0.44
	R2L	0.09	0.34	0.09	0.34	0.06	0.04	0.62	0.12
	DOS	0.00	0.50	0.00	0.50	0.00	0.01	0.75	0.25
	PROBE	0.10	0.88	0.10	0.88	0.10	0.88	0.89	0.39
	U2R	0.01	0.46	0.01	0.46	0.00	0.00	0.73	0.23
						Mean:		**0.79**	**0.29**
OVA boosted See5	NORM	0.24	0.99	0.24	0.99	0.15	0.93	0.91	0.41
	R2L	0.02	0.24	0.02	0.24	0.00	0.01	0.61	0.11
	DOS	0.06	0.56	0.06	0.56	0.01	0.56	0.77	0.27
	PROBE	0.08	0.89	0.08	0.89	0.08	0.89	0.91	0.41
	U2R	0.01	0.40	0.01	0.40	0.00	0.00	0.69	0.19
						Mean:		**0.78**	**0.28**
pVn See5	NORM	0.20	0.98	0.20	0.98	0.07	0.41	0.90	0.40
	R2L	0.02	0.54	0.02	0.54	0.01	0.22	0.76	0.26
	DOS	0.00	0.68	0.00	0.68	0.00	0.44	0.84	0.34
	PROBE	0.03	0.97	0.03	0.97	0.01	0.97	0.98	0.48
	U2R	0.01	0.77	0.01	0.71	0.00	0.43	0.88	0.38
						Mean:		**0.87**	**0.37**

Table G.7: One-vs-rest AUC for the See5 Wine quality models

See5 Wine quality white: TPRATE, FPRATE,auc and MEAN AUC									
		Mean values for thresholds							
		$\lambda_1 = 0.5$		$\lambda_2 = 0.75$		$\lambda_3 = 1.0$			
model	positive Class	FP1	TP1	FP2	TP2	FP3	TP3	*AUC*	AUC_{above}
single See5	4	0.04	0.26	0.04	0.26	0.00	0.01	0.61	0.11
	5	0.33	0.70	0.03	0.05	0.00	0.00	0.68	0.18
	6	0.18	0.28	0.02	0.04	0.00	0.00	0.55	0.05
	7	0.19	0.48	0.05	0.14	0.00	0.00	0.64	0.14
	8	0.01	0.08	0.00	0.08	0.00	0.00	0.54	0.04
						Mean:		**0.60**	**0.10**
un-boosted OVA See5	4	0.09	0.40	0.09	0.40	0.01	0.09	0.66	0.16
	5	0.30	0.65	0.30	0.65	0.02	0.01	0.67	0.17
	6	0.12	0.13	0.10	0.13	0.01	0.00	0.51	0.01
	7	0.25	0.43	0.24	0.43	0.00	0.00	0.59	0.09
	8	0.03	0.13	0.03	0.13	0.00	0.00	0.55	0.05
						Mean:		**0.60**	**0.10**
boosted OVA See5	4	0.09	0.40	0.09	0.40	0.01	0.09	0.66	0.16
	5	0.33	0.68	0.31	0.68	0.02	0.01	0.68	0.18
	6	0.07	0.10	0.01	0.02	0.00	0.00	0.51	0.01
	7	0.26	0.46	0.24	0.45	0.00	0.00	0.60	0.10
	8	0.03	0.14	0.03	0.13	0.00	0.00	0.56	0.06
						Mean:		**0.60**	**0.10**
pVn See5	4	0.06	0.34	0.06	0.34	0.01	0.09	0.64	0.14
	5	0.19	0.55	0.14	0.48	0.00	0.00	0.69	0.19
	6	0.19	0.41	0.12	0.27	0.01	0.02	0.61	0.11
	7	0.29	0.58	0.25	0.56	0.03	0.06	0.66	0.16
	8	0.02	0.14	0.02	0.14	0.01	0.00	0.56	0.06
						Mean:		**0.63**	**0.13**

Appendix H

Using statistical and database software to implement dataset selection methods

Recommendations for using database and statistical software for the implementation of dataset selection methods proposed in this thesis were given in chapter 10. Tables H.1 and H.2 provide detailed suggestions for feature selection, training instance selection and model aggregation.

TableH.1: Suggestions for feature selection using statistical software

Feature selection activity	Step for activity	Implementation
Feature ranking	Generation of probe variables	SPSS, SAS or MS Excel
	Sampling	SPSS or SAS
	Binarisation of qualitative features and class variable	SPSS, SAS or MS Excel
	Measurement of class-feature and feature-feature correlations	Bivariate correlation matrix for quantitative variables
		Pearson's chi-square, SU coefficient, phi and Cramer's V statistics
	Computation of mean and 95% CIs of means for correlations	SPSS
	Ranking and feature elimination using probes	SPSS or MS Excel
Feature subset search	Search for best subset	Specialised code e.g. C++ code

Table H.2: Suggestions for OVA and pVn modeling using statistical software

Activity	Implementation
Sampling for training set to create single model	SPSS or SAS
Creation of single model and confusion matrix	SPSS or SAS
Dataset partitioning	SPSS, SAS or SQL
Sampling from partitions to obtain boosted samples for base model creation	SPSS, SAS
Creation of base models	SPSS, SAS or other modelling software
Model aggregation	SPSS, SAS or MS Excel or Specialised code e.g. C++ code

Appendix I

Publications and conference presentations

LUTU, P. E. N. & ENGELBRECHT, A. P. (2006) A Comparative Study of Sample Selection methods for Classification. *South African Computer Journal,* 36, 69-85.

LUTU, P. E. N. & ENGELBRECHT, A. P. (2008) A decision rule-based method for feature selection in predictive data mining. Presentation at*: The 18th Triennial Conference of the International Federation of Operational Research Societies (IFORS 2008), Sandton, Johannesburg, July 2008*.

LUTU, P. E. N. & ENGELBRECHT, A. P. (2010) A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications,* 37, 602-609.

LUTU, P. E. N. & ENGELBRECHT, A. P. (2010) Using OVA modeling to improve classification performance for large datasets. Submitted to the Journal of Expert Systems With Applications (ESWA).

LUTU, P. E. N. & ENGELBRECHT, A. P. (2010) An algorithm for combining K-Nearest Neighbour base model predictions. Submitted to the Journal of Expert Systems With Applications (ESWA).