Bibliography

Acuna, E., , members of the CASTLE group at UPR-Mayaguez, and Rico., P. (2009). *dprep: Data preprocessing and visualization functions for classification*. R package version 2.1.

Adler, D., Glaser, C., Nenadic, O., Oehlschlagel, J., and Zucchini, W. (2010). *ff: memory-efficient storage of large data on disk and fast access functions*. R package version 2.1-2.

Aha, D. (1990). *A study of instance-based learning algorithms for supervised learning tasks: Mathematical, empirical, and psychological evaluations*. PhD thesis, University of California at Irvine, Department of Information and Computer Science.

Aha, D. (1997). Lazy learning. *Artificial Intelligence Review*, 11.

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Atkeson, C. G., Moore, A., and Schaal, S. (1997). Locally weighted learning. *Artificial Intelligence Review*, 11:11–73.

Austin, V. H. . J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data (3rd edition)*. John Wiley.

Bontempi, G., Birattari, M., and Bersini, H. (1999). Lazy learners at work: the lazy learning toolbox. In *Proceedings of the 7th European Congress on Intelligent Tecnhiques & Soft Computing (EUFIT'99)*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, 26:801–849.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software.

Breunig, M., Kriegel, H., Ng, R., and Sander, J. (2000). Lof: identifying density-based local outliers. In *ACM Int. Conf. on Management of Data*, pages 93–104.

Carl, P. and Peterson, B. G. (2009). *PerformanceAnalytics: Econometric tools for performance and risk analysis.* R package version 1.0.0.

Chambers, J. (2008). *Software for Data Analysis: programming with R.* Springer.

Chan, R. (1999). Protecting rivers & streams by monitoring chemical concentrations and algae communities. In *Proceedings of the 7th European Congress on Intelligent Tecnhiques & Soft Computing (EUFIT'99).*

Chandola, V., Banerjee, A., and Kumar, V. (2007). Outlier detection: a survey. Technical Report TR 07-017, Department of Computer Science and Engineering, University of Minnesota.

Chatfield, C. (1983). *Statistics for technology.* Chapman and Hall, 3rd edition.

Chawla, N. (2005). *The Data Mining and Knowledge Discovery Handbook*, chapter Data mining for imbalanced datasets: An overview, pages 853–867. Springer.

Chawla, N., Japkowicz, N., and Kokz, A. (2004). Sigkdd special issue on learning from imbalanced datasets.

Chawla, N., Lazarevic, A., Hall, L., and Bowyer, K. (2003). Smote-boost: Improving prediction of the minority class in boosting. In *Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 107–119.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321âĂŞ–357.

Chen, C., Hardle, W., and Unwin, A., editors (2008). *Handbook of Data Visualization.* Springer.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7).

Chizi, B. and Maimon, O. (2005). *The Data Mining and Knowledge Discovery Handbook*, chapter Dimension Reduction and Feature Selection, pages 93–111. Springer.

Cleveland, W. (1993). *Visualizing Data.* Hobart Press.

Cleveland, W. (1995). *The Elements of Graphing Data*. Hobart Press.

Cleveland, W. and Loader, C. (1995). Smoothing by local regression: Principles and methods (with discussion). *Computational Statistics*.

Cortes, E. A., Martinez, M. G., and Rubio, N. G. (2010). *adabag: Applies Adaboost.M1 and Bagging*. R package version 1.1.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Dalgaard, P. (2002). *Introductory Statistics with R*. Springer.

Deboeck, G., editor (1994). *Trading on the edge*. John Wiley & Sons.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Devogelaere, D., Rijckaert, M., and Embrechts, M. J. (1999). 3rd international competition: Protecting rivers and streams by monitoring chemical concentrations and algae communities solved with the use of gadc. In *Proceedings of the 7th European Congress on Intelligent Tecnhiques & Soft Computing (EUFIT'99)*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., , and Weingessel, A. (2009). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-19.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *KDD'99: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, pages 155âĂŞ–164. ACM Press.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–137.

Drapper, N. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley & Sons, 2nd edition.

Drummond, C. and Holte, R. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130.

DuBois, P. (2000). *MySQL*. New Riders.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *IJCAI'01: Proceedings of 17th International Joint Conference of Artificial Intelligence*, pages 973–978. Morgan Kaufmann Publishers Inc.

Fox, J. (2009). *car: Companion to Applied Regression*. R package version 1.2-16.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*.

Freund, Y. and Shapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*. Morgan Kaufmann.

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–144.

Friedman, J. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378.

Gama, J. and Gaber, M., editors (2007). *Learning from data streams*. Springer.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with drift detection. In Bazzan, A. and Labidi, S., editors, *Advances in Artificial Intelligence-SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pages 286–295. Springer.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2010). *genefilter: genefilter: methods for filtering genes from microarray experiments*. R package version 1.28.2.

Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Hahne, F., Huber, W., Gentleman, R., and Falcon, S. (2008). *Bioconductor Case Studies*. Springer.

Han, J. and Kamber, M. (2006). *Data Mining: concepts and techniques (2nd edition)*. Morgan Kaufmann Publishers.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Hand, D. and Yu, K. (2001). Idiot's bayes - not so stupid after all? *International Statistical Review*, 69(3):385–399.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123.

Harrell Jr, F. E. (2009). *Hmisc: Harrell Miscellaneous*. R package version 3.7-0. With contributions from many other users.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction.* Springer.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman and Hall.

Hong, S. (1997). Use of contextual information for feature ranking and discretization. *IEEE Transactions on Knowledge and Data Engineering.*

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

James, D. A. and DebRoy, S. (2009). *RMySQL: R interface to the MySQL database.* R package version 0.7-4.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-A1'2000):Special Track on Inductive Learning.*

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *VLDB 04: Proceedings of the 30th International Conference on Very Large Data Bases*, pages 180–191. Morgan Kaufmann.

Kira, K. and Rendel, L. (1992). The feature selection problem : Traditional methods and a new algorithm. In *Proc. Tenth National Conference on Artificial Intelligence*, pages 129–134. MIT Press.

Klinkenberg, R. (2004). Learning drifting concepts: example selection vs. example weighting. *Intelligent Data Analysis*, 8(3):281–300.

Kononenko, I. (1991). Semi-naive bayesian classifier. In *EWSL-91: Proceedings of the European working session on learning on Machine learning*, pages 206–219. Springer-Verlag New York, Inc.

Kononenko, I., Simec, E., and Robnik-Sikonja, M. (1997). Overcoming the myopia of induction learning algorithms with relieff. *Applied Intelligence*, 17(1):39–55.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186.

Leisch, F., Hornik, K., and Ripley., B. D. (2009). *mda: Mixture and flexible discriminant analysis, S original by Trevor Hastie and Robert Tibshirani*. R package version 0.3-4.

Li, X. (2009). *ALL: A data package*. R package version 1.4.7.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Milborrow, S. (2009). *earth: Multivariate Adaptive Regression Spline Models, derived from mda:mars by Trevor Hastie and Rob Tibshirani*. R package version 2.4-0.

Minsky, M. and Papert, S. (1969). *Perceptrons: an introduction to computational geometry*. MIT Press.

Murrell, P. (2006). *R Graphics*. Chapman & Hall/CRC.

Murtagh, F. (1985). Multidimensional clustering algorithms. *COMPSTAT Lectures 4, Wuerzburg: Physica-Verlag*.

Myers, R. (1990). *Classical and modern regression with applications*. Duxbury Press, 2nd edition.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Nemenyi, P. (1969). *Distribution-free multiple comparisons*. PhD thesis, Princeton University.

Ng, R. and Han, J. (1994). Efficient and efective clustering method for spatial data mining. In *Proc. of VLDB'94*.

Oakland, J. (2007). *Statistical Process Control, 6th edition*. Butterworth-Heinemann.

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD'97: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3).

Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proc. 15th International Conf. on Machine Learning*, pages 445–453. Morgan Kaufmann, San Francisco, CA.

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.

Quinlan, R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers.

R Special Interest Group on Databases, R.-S.-D. (2009). *DBI: R Database Interface*. R package version 0.2-5.

Rätsch, G., Onoda, T., and Müller, K. (2001). Soft margins for adaboost. *Machine Learning*, 42(3):287–320.

Rijsbergen, C. V. (1979). *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition.

Rish, I. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.

Rogers, R. and Vemuri, V. (1994). *Artificial neural networks forecasting time series*. IEEE Computer Society Press.

Rojas, R. (1996). *Neural Networks*. Springer-Verlag.

Ronsenblatt, F. (1958). The perceptron: a probabilistic models for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *Proc. of the 7th IEEE Workshop on Applications of Computer Vision*, pages 29–36. IEEE Computer Society.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In et. al., D. R., editor, *Parallel distributed processing*, volume 1. MIT Press.

Ryan, J. A. (2009). *quantmod: Quantitative Financial Modelling Framework*. R package version 0.3-13.

Ryan, J. A. and Ulrich, J. M. (2010). *xts: Extensible Time Series*. R package version 0.7-0.

Sarkar, D. (2010). *lattice: Lattice Graphics*. R package version 0.18-3.

Seeger, M. (2002). Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh.

Shapire, R. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.

Shawe-Taylor, J. and Cristianini, N. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2009). *ROCR: Visualizing the performance of scoring classifiers*. R package version 1.0-4.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.*, 14:199–222.

Smola, A. J. and Schölkopf, B. (1998). A tutorial on support vector regression. In *NeuroCOLT Technical Report TR-98-030*.

Therneau, T. M. and port by Brian Ripley., B. A. R. (2010). *rpart: Recursive Partitioning*. R package version 3.1-46.

Torgo, L. (1999a). *Inductive Learning of Tree-based Regression Models*. PhD thesis, Faculty of Sciences, University of Porto.

Torgo, L. (1999b). Predicting the density of algae communities using local regression trees. In *Proceedings of the 7th European Congress on Intelligent Tecnhiques & Soft Computing (EUFIT'99)*.

Torgo, L. (2000). Partial linear trees. In Langley, P., editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1007–1014. Morgan Kaufmann Publishers.

Torgo, L. (2007). Resource-bounded fraud detection. In et. al, N., editor, *Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA'07)*, LNAI. Springer.

Trapletti, A. and Hornik, K. (2009). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-22.

Ulrich, J. (2009). *TTR: Technical Trading Rules*. R package version 0.20-1.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26:359–372.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klar analyzing german business cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

Weiss, G. and F.Provost (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354.

Weiss, S. and Indurkhya, N. (1999). *Predictive data mining*. Morgan Kaufmann.

Werbos, P. (1974). *Beyond regression - new tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University.

Werbos, P. (1996). *The roots of backpropagation - from observed derivatives to neural networks and political forecasting*. John Wiley & Sons.

Wettschereck, D. (1994). *A study of distance-based machine learning algorithms*. PhD thesis, Oregon State University.

Wettschereck, D., Aha, D., and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazzy learning algorithms. *Artificial Intelligence Review*, 11:11–73.

Wilson, D. and Martinez, T. (1997). Improved heterogeneous distance functions. *JAIR*, 6:1–34.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

Zhu, X. (2005). *Semi-supervised learning with graphs*. PhD thesis, School of Computer Science, Carnegie Mellon University.

Zhu, X. (2006). Semi-supervised literature survey. Technical Report TR 1530, University of Wisconsin-Madison.

Zirilli, J. (1997). *Financial prediction using neural networks*. Thomson Computer Press.