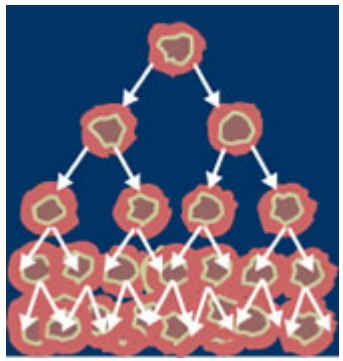
Center for Machine Learning and Intelligent Systems

Breast Cancer Data Set

Download: Data Folder, Data Set Description

Abstract: Breast Cancer Data (Restricted Access)



Data Set Characteristics:	Multivariate	**Number of Instances:**	286	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	9	**Date Donated**	1988-07-11
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	313291

Source:

Creators:

Matjaz Zwitter & Milan Soklic (physicians)
Institute of Oncology
University Medical Center
Ljubljana, Yugoslavia

Donors:

Ming Tan and Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

Data Set Information:

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Attribute Information:

1. Class: no-recurrence-events, recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: yes, no.
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. irradiat: yes, no.

Relevant Papers:

Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.
[Web Link]

Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press.
[Web Link]

Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.
[Web Link]

Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.
[Web Link]

Papers That Cite This Data Set[1]:

Igor Fischer and Jan Poland. Amplifying the Block Matrix Structure for Spectral Clustering. Telecommunications Lab. 2005. [View Context].

Saher Esmeir and Shaul Markovitch. Lookahead-based algorithms for anytime induction of decision trees. ICML. 2004. [View Context].

Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004. [View Context].

Kaizhu Huang and Haiqin Yang and Irwin King and Michael R. Lyu and Laiwan Chan. Biased Minimax Probability Machine for Medical Diagnosis. AMAI. 2004. [View Context].

Qingping Tao Ph. D. MAKING EFFICIENT LEARNING ALGORITHMS WITH EXPONENTIALLY MANY FEATURES. Qingping Tao A DISSERTATION Faculty of The Graduate College University of Nebraska In Partial Fulfillment of Requirements. 2004. [View Context].

Krzysztof Grabczewski and Wl/odzisl/aw Duch. Heterogeneous Forests of Decision Trees. ICANN. 2002. [View Context].

Hussein A. Abbass. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine, 25. 2002. [View Context].

Fei Sha and Lawrence K. Saul and Daniel D. Lee. Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines. NIPS. 2002. [View Context].

Kristin P. Bennett and Ayhan Demiriz and Richard Maclin. Exploiting unlabeled data in ensemble methods. KDD. 2002. [View Context].

Baback Moghaddam and Gregory Shakhnarovich. Boosted Dyadic Kernel Discriminants. NIPS. 2002. [View Context].

András Antos and Balázs Kégl and Tamás Linder and Gábor Lugosi. Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research, 3. 2002. [View Context].

Michael G. Madden. Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm. CoRR, csLG/0211003. 2002. [View Context].

Yongmei Wang and Ian H. Witten. Modeling for Optimal Probability Prediction. ICML. 2002. [View Context].

Remco R. Bouckaert. Accuracy bounds for ensembles under 0 { 1 loss. Xtal Mountain Information Technology & Computer Science Department, University of Waikato. 2002. [View Context].

Nikunj C. Oza and Stuart J. Russell. Experimental comparisons of online and batch versions of bagging and boosting. KDD. 2001. [View Context].

Bernhard Pfahringer and Geoffrey Holmes and Richard Kirkby. Optimizing the Induction of Alternating Decision Trees. PAKDD. 2001. [View Context].

Robert Burbidge and Matthew Trotter and Bernard F. Buxton and Sean B. Holden. STAR - Sparsity through Automated Rejection. IWANN (1). 2001. [View Context].

Bernhard Pfahringer and Geoffrey Holmes and Gabi Schmidberger. Wrapping Boosters against Noise. Australian Joint Conference on Artificial Intelligence. 2001. [View Context].

W. Nick Street and Yoo-Hyon Kim. A streaming ensemble algorithm (SEA) for large-scale classification. KDD. 2001. [View Context].

Lorne Mason and Peter L. Bartlett and Jonathan Baxter. Improved Generalization Through Explicit Optimization of Margins. Machine Learning, 38. 2000. [View Context].

Endre Boros and Peter Hammer and Toshihide Ibaraki and Alexander Kogan and Eddy Mayoraz and Ilya B. Muchnik. An Implementation of Logical Analysis of Data. IEEE Trans. Knowl. Data Eng, 12. 2000. [View Context].

P. S and Bradley K. P and Bennett A. Demiriz. Constrained K-Means Clustering. Microsoft Research Dept. of Mathematical Sciences One Microsoft Way Dept. of Decision Sciences and Eng. Sys. 2000. [View Context].

Sally A. Goldman and Yan Zhou. Enhancing Supervised Learning with Unlabeled Data. ICML. 2000. [View Context].

Justin Bradley and Kristin P. Bennett and Bennett A. Demiriz. Constrained K-Means Clustering. Microsoft Research Dept. of Mathematical Sciences One Microsoft Way Dept. of Decision Sciences and Eng. Sys. 2000. [View Context].

Yuh-Jeng Lee. Smooth Support Vector Machines. Preliminary Thesis Proposal Computer Sciences Department University of Wisconsin. 2000. [View Context].

Petri Kontkanen and Petri Myllym and Tomi Silander and Henry Tirri and Peter Gr. On predictive distributions and Bayesian networks. Department of Computer Science, Stanford University. 2000. [View Context].

Kristin P. Bennett and Ayhan Demiriz and John Shawe-Taylor. A Column Generation Algorithm For Boosting. ICML. 2000. [View Context].

Matthew Mullin and Rahul Sukthankar. Complete Cross-Validation for Nearest Neighbor Classifiers. ICML. 2000. [View Context].

Chun-Nan Hsu and Hilmar Schuschel and Ya-Ting Yang. The ANNIGMA-Wrapper Approach to Neural Nets Feature Selection for Knowledge Discovery and Data Mining. Institute of Information Science. 1999. [View Context].

David M J Tax and Robert P W Duin. Support vector domain description. Pattern Recognition Letters, 20. 1999. [View Context].

Kai Ming Ting and Ian H. Witten. Issues in Stacked Generalization. J. Artif. Intell. Res. (JAIR, 10. 1999. [View Context].

Ismail Taha and Joydeep Ghosh. Symbolic Interpretation of Artificial Neural Networks. IEEE Trans. Knowl. Data Eng, 11. 1999. [View Context].

Lorne Mason and Jonathan Baxter and Peter L. Bartlett and Marcus Frean. Boosting Algorithms as Gradient Descent. NIPS. 1999. [View Context].

Iñaki Inza and Pedro Larrañaga and Basilio Sierra and Ramon Etxeberria and Jose Antonio Lozano and Jos Manuel Peña. Representing the behaviour of supervised classification learning algorithms by Bayesian networks. Pattern Recognition Letters, 20. 1999. [View Context].

David W. Opitz and Richard Maclin. Popular Ensemble Methods: An Empirical Study. J. Artif. Intell. Res. (JAIR, 11. 1999. [View Context].

Lorne Mason and Peter L. Bartlett and Jonathan Baxter. Direct Optimization of Margins Improves Generalization in Combined Classifiers. NIPS. 1998. [View Context].

Richard Maclin. Boosting Classifiers Regionally. AAAI/IAAI. 1998. [View Context].

Huan Liu and Hiroshi Motoda and Manoranjan Dash. A Monotonic Measure for Optimal Feature Selection. ECML. 1998. [View Context].

Yk Huhtala and Juha Kärkkäinen and Pasi Porkka and Hannu Toivonen. Efficient Discovery of Functional and Approximate Dependencies Using Partitions. ICDE. 1998. [View Context].

W. Nick Street. A Neural Network Model for Prognostic Prediction. ICML. 1998. [View Context].

Kristin P. Bennett and Erin J. Bredensteiner. A Parametric Optimization Method for Machine Learning. INFORMS Journal on Computing, 9. 1997. [View Context].

Pedro Domingos. Control-Sensitive Feature Selection for Lazy Learners. Artif. Intell. Rev, 11. 1997. [View Context].

Rudy Setiono and Huan Liu. NeuroLinear: From neural networks to oblique decision rules. Neurocomputing, 17. 1997. [View Context].

. Prototype Selection for Composite Nearest Neighbor Classifiers. Department of Computer Science University of Massachusetts. 1997. [View Context].

Ismail Taha and Joydeep Ghosh. Characterization of the Wisconsin Breast cancer Database Using a Hybrid Symbolic-Connectionist System. Proceedings of ANNIE. 1996. [View Context].

Kamal Ali and Michael J. Pazzani. Error Reduction through Learning Multiple Descriptions. Machine Learning, 24. 1996. [View Context].

Jennifer A. Blue and Kristin P. Bennett. Hybrid Extreme Point Tabu Search. Department of Mathematical Sciences Rensselaer Polytechnic Institute. 1996. [View Context].

Pedro Domingos. Unifying Instance-Based and Rule-Based Induction. Machine Learning, 24. 1996. [View Context].

Erin J. Bredensteiner and Kristin P. Bennett. Feature Minimization within Decision Trees. National Science Foundation. 1996. [View Context].

Geoffrey I. Webb. OPUS: An Efficient Admissible Algorithm for Unordered Search. J. Artif. Intell. Res. (JAIR, 3. 1995. [View Context].

Christophe Giraud and Tony Martinez and Christophe G. Giraud-Carrier. University of Bristol Department of Computer Science ILA: Combining Inductive Learning with Prior Knowledge and Reasoning. 1995. [View Context].

Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI. 1995. [View Context].

Rong-En Fan and P. -H Chen and C. -J Lin. Working Set Selection Using the Second Order Information for Training SVM. Department of Computer Science and Information Engineering National Taiwan University. [View Context].

Rong Jin and Yan Liu and Luo Si and Jaime Carbonell and Alexander G. Hauptmann. A New Boosting Algorithm Using Input-Dependent Regularizer. School of Computer Science, Carnegie Mellon University. [View Context].

David Kwartowitz and Sean Brophy and Horace Mann. Session S2D Work In Progress: Establishing multiple contexts for student's progressive refinement of data mining. [View Context].

Geoffrey I Webb. Generality is more significant than complexity: Toward an alternative to Occam's Razor. School of Computing and Mathematics Deakin University. [View Context].

Karthik Ramakrishnan. UNIVERSITY OF MINNESOTA. [View Context].

Geoffrey I Webb. Learning Decision Lists by Prepending Inferred Rules. School of Computing and Mathematics Deakin University. [View Context].

Adil M. Bagirov and Alex Rubinov and A. N. Soukhojak and John Yearwood. Unsupervised and supervised data classification via nonsmooth and global optimization. School of Information Technology and Mathematical Sciences, The University of Ballarat. [View Context].

M. V. Fidelis and Heitor S. Lopes and Alex Alves Freitas. Discovering Comprehensible Classification Rules with a Genetic Algorithm. UEPG, CPD CEFET-PR, CPGEI PUC-PR, PPGIA Praa Santos Andrade, s/n Av. Sete de Setembro. [View Context].

Chris Drummond and Robert C. Holte. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. Institute for Information Technology, National Research Council Canada. [View Context].

Wl odzisl/aw Duch and Rudy Setiono and Jacek M. Zurada. Computational intelligence methods for rule-based data understanding. [View Context].

Maria Salamo and Elisabet Golobardes. Analysing Rough Sets weighting methods for Case-Based Reasoning Systems. Enginyeria i Arquitectura La Salle. [View Context].

G. Ratsch and B. Scholkopf and Alex Smola and K. -R Muller and T. Onoda and Sebastian Mika. Arc: Ensemble Learning in the Presence of Outliers. GMD FIRST. [View Context].

D. Randall Wilson and Roel Martinez. Improved Center Point Selection for Probabilistic Neural Networks. Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms. [View Context].

Chiranjib Bhattacharyya. Robust Classification of noisy data using Second Order Cone Programming approach. Dept. Computer Science and Automation, Indian Institute of Science. [View Context].

K. A. J Doherty and Rolf Adams and Neil Davey. Unsupervised Learning with Normalised Data and Non-Euclidean Norms. University of Hertfordshire. [View Context].

Adam H. Cannon and Lenore J. Cowen and Carey E. Priebe. Approximate Distance Classification. Department of Mathematical Sciences The Johns Hopkins University. [View Context].

G. Ratsch and B. Scholkopf and Alex Smola and Sebastian Mika and T. Onoda and K. -R Muller. Robust Ensemble Learning for Data Mining. GMD FIRST, Kekul#estr. [View Context].

Andrew I. Schein and Lyle H. Ungar. A-Optimality for Active Learning of Logistic Regression Classifiers. Department of Computer and Information Science Levine Hall. [View Context].

Huan Liu. A Family of Efficient Rule Generators. Department of Information Systems and Computer Science National University of Singapore. [View Context].

Alexander K. Seewald. Dissertation Towards Understanding Stacking Studies of a General Ensemble Learning Scheme ausgefuhrt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Naturwissenschaften. [View Context].

Rafael S. Parpinelli and Heitor S. Lopes and Alex Alves Freitas. PART FOUR: ANT COLONY OPTIMIZATION AND IMMUNE SYSTEMS Chapter X An Ant Colony Algorithm for Classification Rule Discovery. CEFET-PR, Curitiba. [View Context].

Paul D. Wilson and Tony R. Martinez. Combining Cross-Validation and Confidence to Measure Fitness. fonix corporation Brigham Young University. [View Context].

Charles Campbell and Nello Cristianini. Simple Learning Algorithms for Training Support Vector Machines. Dept. of Engineering Mathematics. [View Context].

Nikunj C. Oza and Stuart J. Russell. Online Bagging and Boosting. Computer Science Division University of California. [View Context].

Michael R. Berthold and Klaus--Peter Huber. From Radial to Rectangular Basis Functions: A new Approach for Rule Learning from Large Datasets. Institut fur Rechnerentwurf und Fehlertoleranz (Prof. D. Schmid) Universitat Karlsruhe. [View Context].

Bart Baesens and Stijn Viaene and Tony Van Gestel and J. A. K Suykens and Guido Dedene and Bart De Moor and Jan Vanthienen and Katholieke Universiteit Leuven. An

Empirical Assessment of Kernel Type Performance for Least Squares Support Vector Machine Classifiers. Dept. Applied Economic Sciences. [View Context].

Rudy Setiono and Huan Liu. Neural-Network Feature Selector. Department of Information Systems and Computer Science National University of Singapore. [View Context].

Rafael S. Parpinelli and Heitor S. Lopes and Alex Alves Freitas. An Ant Colony Based System for Data Mining: Applications to Medical Data. CEFET-PR, CPGEI Av. Sete de Setembro, 3165. [View Context].

Wl odzisl and Rafal Adamczak and Krzysztof Grabczewski and Grzegorz Zal. A hybrid method for extraction of logical rules from data. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Jarkko Salojarvi and Samuel Kaski and Janne Sinkkonen. Discriminative clustering in Fisher metrics. Neural Networks Research Centre Helsinki University of Technology. [View Context].

Rudy Setiono. Extracting M-of-N Rules from Trained Neural Networks. School of Computing National University of Singapore. [View Context].

Ayhan Demiriz and Kristin P. Bennett and John Shawe and I. Nouretdinov V.. Linear Programming Boosting via Column Generation. Dept. of Decision Sciences and Eng. Systems, Rensselaer Polytechnic Institute. [View Context].

Liping Wei and Russ B. Altman. An Automated System for Generating Comparative Disease Profiles and Making Diagnoses. Section on Medical Informatics Stanford University School of Medicine, MSOB X215. [View Context].

Chotirat Ann and Dimitrios Gunopulos. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Computer Science Department University of California. [View Context].

Sherrie L. W and Zijian Zheng. A BENCHMARK FOR CLASSIFIER LEARNING. Basser Department of Computer Science The University of Sydney. [View Context].

John W. Chinneck. Fast Heuristics for the Maximum Feasible Subsystem Problem. Systems and Computer Engineering, Carleton University. [View Context].

M. A. Galway and Michael G. Madden. DEPARTMENT OF INFORMATION TECHNOLOGY technical report NUIG-IT-011002 Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm. Department of Information Technology National University of Ireland, Galway. [View Context].

John G. Cleary and Leonard E. Trigg. Experiences with OB1, An Optimal Bayes Decision Tree Learner. Department of Computer Science University of Waikato. [View Context].

Wl/odzisl/aw Duch and Rafal/ Adamczak Email:duchraad@phys. uni. torun. pl. Statistical methods for construction of neural networks. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Citation Request:

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. Please include this citation if you plan to use this database.

---

[1] Papers were automatically harvested and associated with this data set, in collaboration with Rexa.info