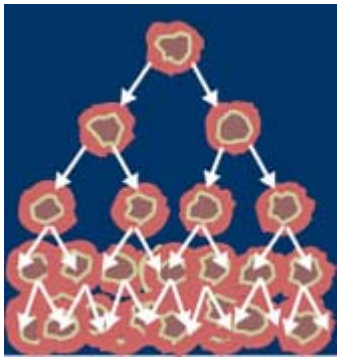
Breast Cancer Wisconsin (Diagnostic) Data Set

Download: Data Folder, Data Set Description

Abstract: Diagnostic Wisconsin Breast Cancer Database



Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	17533

Source:

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.
University of Wisconsin, Clinical Sciences Center
Madison, WI 53792
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi '@' cs.wisc.edu

Donor:

Nick Street

Data Set Information:

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at [Web Link]

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Relevant Papers:

First Usage:

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.
[Web Link]

O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.
[Web Link]

Medical literature:

W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.
[Web Link]

W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pages 77-87, April 1995.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery 1995;130:511-516.
[Web Link]

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792--796, 1995.
[Web Link]

See also:

[Web Link]
[Web Link]

Papers That Cite This Data Set[1]:

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

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

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

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

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

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

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

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

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].

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].

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].

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

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

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

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

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

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

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

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

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].

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

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

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

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

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

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].

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

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].

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].

Citation Request:

Please refer to the Machine Learning Repository's citation policy

---

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