Breast Cancer Wisconsin (Original) Data Set
Download: Data Folder, Data Set Description

Abstract: Original Wisconsin Breast Cancer Database

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	409278

Source:

Creator:

Dr. WIlliam H. Wolberg (physician)
University of Wisconsin Hospitals
Madison, Wisconsin, USA

Donor:

Olvi Mangasarian (mangasarian '@' cs.wisc.edu)
Received by David W. Aha (aha '@' cs.jhu.edu)

Data Set Information:

Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself:

Group 1: 367 instances (January 1989)

Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated datbase on 15 July 1992)

Note that the results summarized above in Past Usage refer to a dataset of size 369, while Group 1 has only 367 instances. This is because it originally contained 369 instances; 2 were removed. The following statements summarizes changes to the original Group 1's set of data:

Group 1 : 367 points: 200B 167M (January 1989)

Revised Jan 10, 1991: Replaced zero bare nuclei in 1080185 & 1187805

Revised Nov 22,1991: Removed 765878,4,5,9,7,10,10,10,3,8,1 no record
: Removed 484201,2,7,8,8,4,3,10,3,4,1 zero epithelial
: Changed 0 to 1 in field 6 of sample 1219406
: Changed 0 to 1 in field 8 of following sample:
: 1182404,2,3,1,1,1,2,0,1,1,1

Attribute Information:

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

Relevant Papers:

Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.
[Web Link]

Zhang, J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470--479). Aberdeen, Scotland: Morgan Kaufmann.

[Web Link]

Papers That Cite This Data Set[1]:

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

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].

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

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

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

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

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].

W. Nick Street. A Neural Network Model for Prognostic Prediction. ICML. 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].

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

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].

. Prototype Selection for Composite Nearest Neighbor Classifiers. Department of Computer Science University of Massachusetts. 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].

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

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].

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

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

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].

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

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].

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].

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

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

Citation Request:

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. If you publish results when using this database, then please include this information in your acknowledgements. Also, please cite one or more of:
1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
2. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical

diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
3. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
4. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).
