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).

1. Title: Wisconsin Breast Cancer Database (January 8, 1991)

2. Sources:

-- 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)
-- Date: 15 July 1992

3. Past Usage:

Attributes 2 through 10 have been used to represent instances.
Each instance has one of 2 possible classes: benign or malignant.

1. Wolberg,~W.~H., \& Mangasarian,~O.~L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In {\it Proceedings of the National Academy of Sciences}, {\it 87}, 9193--9196.
   -- Size of data set: only 369 instances (at that point in time)
   -- Collected classification results: 1 trial only
   -- Two pairs of parallel hyperplanes were found to be consistent with 50% of the data
      -- Accuracy on remaining 50% of dataset: 93.5%
   -- Three pairs of parallel hyperplanes were found to be consistent with 67% of data
      -- Accuracy on remaining 33% of dataset: 95.9%

2. Zhang,~J. (1992). Selecting typical instances in instance-based learning. In {\it Proceedings of the Ninth International Machine Learning Conference} (pp. 470--479). Aberdeen, Scotland: Morgan Kaufmann.
   -- Size of data set: only 369 instances (at that point in time)
   -- Applied 4 instance-based learning algorithms
   -- Collected classification results averaged over 10 trials
   -- Best accuracy result:
      -- 1-nearest neighbor: 93.7%
      -- trained on 200 instances, tested on the other 169
   -- Also of interest:
      -- Using only typical instances: 92.2% (storing only 23.1 instances)
      -- trained on 200 instances, tested on the other 169

4. Relevant 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

5. Number of Instances: 699 (as of 15 July 1992)

6. Number of Attributes: 10 plus the class atribute (11)

7. Attribute Information: (class attribute has been moved to last column)

Attribute	Domain
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

8. Missing attribute values: 16

There are 16 instances in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".

9. Class distribution:

Benign: 458 (65.5%)

Malignant: 241 (34.5%)