1. Title: Wisconsin Prognostic Breast Cancer (WPBC)

2. Source Information

a) Creators:

Dr. William H. Wolberg, General Surgery Dept., University of

Wisconsin, Clinical Sciences Center, Madison, WI 53792

wolberg@eagle.surgery.wisc.edu

W. Nick Street, Computer Sciences Dept., University of

Wisconsin, 1210 West Dayton St., Madison, WI 53706

street@cs.wisc.edu 608-262-6619

Olvi L. Mangasarian, Computer Sciences Dept., University of

Wisconsin, 1210 West Dayton St., Madison, WI 53706

olvi@cs.wisc.edu

b) Donor: Nick Street

c) Date: December 1995

3. Past Usage:

Various versions of this data have been used in the following

publications:

(i) W. N. Street, O. L. Mangasarian, and W.H. Wolberg.

An inductive learning approach to prognostic prediction.

In A. Prieditis and S. Russell, editors, Proceedings of the

Twelfth International Conference on Machine Learning, pages

522--530, San Francisco, 1995. Morgan Kaufmann.

(ii) 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.

(iii) 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.

(iv) 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.

(v) W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian.

Computer-derived nuclear ``grade'' and breast cancer prognosis.

Analytical and Quantitative Cytology and Histology, Vol. 17,

pages 257-264, 1995.

See also:

http://www.cs.wisc.edu/~olvi/uwmp/mpml.html

http://www.cs.wisc.edu/~olvi/uwmp/cancer.html

Results:

Two possible learning problems:

1) Predicting field 2, outcome: R = recurrent, N = nonrecurrent

- Dataset should first be filtered to reflect a particular

endpoint; e.g., recurrences before 24 months = positive,

nonrecurrence beyond 24 months = negative.

- 86.3% accuracy estimated accuracy on 2-year recurrence using

previous version of this data. Learning method: MSM-T (see

below) in the 4-dimensional space of Mean Texture, Worst Area,

Worst Concavity, Worst Fractal Dimension.

2) Predicting Time To Recur (field 3 in recurrent records)

- Estimated mean error 13.9 months using Recurrence Surface

Approximation. (See references (i) and (ii) above)

4. Relevant information

Each record represents follow-up data for one breast cancer

case. These are consecutive patients seen by Dr. Wolberg

since 1984, and include only those cases exhibiting invasive

breast cancer and no evidence of distant metastases at the

time of diagnosis.

The first 30 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

http://www.cs.wisc.edu/~street/images/

The separation 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].

The Recurrence Surface Approximation (RSA) method is a linear

programming model which predicts Time To Recur using both

recurrent and nonrecurrent cases. See references (i) and (ii)

above for details of the RSA method.

This database is also available through the UW CS ftp server:

ftp ftp.cs.wisc.edu

cd math-prog/cpo-dataset/machine-learn/WPBC/

5. Number of instances: 198

6. Number of attributes: 34 (ID, outcome, 32 real-valued input features)

7. Attribute information

1) ID number

2) Outcome (R = recur, N = nonrecur)

3) Time (recurrence time if field 2 = R, disease-free time if

field 2 = N)

4-33) 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)

Several of the papers listed above contain detailed descriptions of

how these features are computed.

The mean, standard error, and "worst" or largest (mean of the three

largest values) of these features were computed for each image,

resulting in 30 features. For instance, field 4 is Mean Radius, field

14 is Radius SE, field 24 is Worst Radius.

Values for features 4-33 are recoded with four significant digits.

34) Tumor size - diameter of the excised tumor in centimeters

35) Lymph node status - number of positive axillary lymph nodes

observed at time of surgery

8. Missing attribute values:

Lymph node status is missing in 4 cases.

9. Class distribution: 151 nonrecur, 47 recur