Wine Classification Problem and Data Description

1. Title of Database: Wine recognition data
Updated Sept 21, 1998 by C.Blake : Added attribute information

2. Sources:
(a) Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

(b) Stefan Aeberhard, email: stefan@coral.cs.jcu.edu.au
(c) July 1991

3. Past Usage:

(1)
S. Aeberhard, D. Coomans and O. de Vel,
Comparison of Classifiers in High Dimensional Settings,
Tech. Rep. no. 92-02, (1992), Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.
(Also submitted to Technometrics).

The data was used with many others for comparing various classifiers. The classes are separable, though only RDA has achieved 100% correct classification.
(RDA : 100%, QDA 99.4%, LDA 98.9%, 1NN 96.1% (z-transformed data))
(All results using the leave-one-out technique)

In a classification context, this is a well posed problem with "well behaved" class structures. A good data set for first testing of a new classifier, but not very challenging.

(2)
S. Aeberhard, D. Coomans and O. de Vel,
"THE CLASSIFICATION PERFORMANCE OF RDA"
Tech. Rep. no. 92-01, (1992), Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.
(Also submitted to Journal of Chemometrics).

Here, the data was used to illustrate the superior performance of the use of a new appreciation function with RDA.

4. Relevant Information:

-- These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.
The analysis determined the quantities of 13 constituents found in each of the three types of wines.

```
   -- I think that the initial data set had around 30 variables, but
      for some reason I only have the 13 dimensional version.
      I had a list of what the 30 or so variables were, but a.)
      I lost it, and b.), I would not know which 13 variables
      are included in the set.

   -- The attributes are (dontated by Riccardo Leardi,
        riclea@anchem.unige.it )
        1) Alcohol
        2) Malic acid
        3) Ash
        4) Alcalinity of ash
        5) Magnesium
        6) Total phenols
        7) Flavanoids
        8) Nonflavanoid phenols
        9) Proanthocyanins
        10)Color intensity
        11)Hue
        12)OD280/OD315 of diluted wines
        13)Proline

5. Number of Instances

        class 1 59
        class 2 71
        class 3 48

6. Number of Attributes

        13

7. For Each Attribute:

        All attributes are continuous

        No statistics available, but suggest to standardise
        variables for certain uses (e.g. for us with classifiers
        which are NOT scale invariant)

        NOTE: 1st attribute is class identifier (1-3)

8. Missing Attribute Values:

        None

9. Class Distribution: number of instances per class

        class 1 59
        class 2 71
        class 3 48


10. Papers That Cite This Data Set:
```

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

Ping Zhong and Masao Fukushima. A Regularized Nonsmooth Newton Method for Multi-class Support Vector Machines. 2005. [View Context].

Stefan Mutter and Mark Hall and Eibe Frank. Using Classification to Evaluate the Output of Confidence-Based Association Rule Mining. Australian Conference on Artificial Intelligence. 2004. [View Context].

Jennifer G. Dy and Carla Brodley. Feature Selection for Unsupervised Learning. Journal of Machine Learning Research, 5. 2004. [View Context].

Yuan Jiang and Zhi-Hua Zhou. Editing Training Data for kNN Classifiers with Neural Network Ensemble. ISNN (1). 2004. [View Context].

Mikhail Bilenko and Sugato Basu and Raymond J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. ICML. 2004. [View Context].

Agapito Ledezma and Ricardo Aler and Araceli Sanchís and Daniel Borrajo. Empirical Evaluation of Optimized Stacking Configurations. ICTAI. 2004. [View Context].

Jianbin Tan and David L. Dowe. MML Inference of Oblique Decision Trees. Australian Conference on Artificial Intelligence. 2004. [View Context].

Sugato Basu. Semi-Supervised Clustering with Limited Background Knowledge. AAAI. 2004. [View Context].

Jeremy Kubica and Andrew Moore. Probabilistic Noise Identification and Data Cleaning. ICDM. 2003. [View Context].

Sugato Basu. Also Appears as Technical Report, UT-AI. PhD Proposal. 2003. [View Context].

Michael L. Raymer and Travis E. Doom and Leslie A. Kuhn and William F. Punch. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 33. 2003. [View Context].

Mukund Deshpande and George Karypis. Using conjunction of attribute values for classification. CIKM. 2002. [View Context].

Petri Kontkanen and Jussi Lahtinen and Petri Myllymaki and Tomi Silander and Henry Tirri. Proceedings of Pre- and Post-processing in Machine Learning and Data Mining: Theoretical Aspects and Applications, a workshop within Machine Learning and Applications. Complex Systems Computation Group (CoSCo). 1999. [View Context].

Ethem Alpaydin. Voting over Multiple Condensed Nearest Neighbors. Artif. Intell. Rev, 11. 1997. [View Context].

Georg Thimm and E. Fiesler. Optimal Setting of Weights, Learning Rate, and Gain. E S E A R C H R E P R O R T I D I A P. 1997. [View Context].

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

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

Georg Thimm and Emile Fiesler. IDIAP Technical report High Order and Multilayer Perceptron Initialization. IEEE Transactions. 1994. [View Context].

Abdelhamid Bouchachia. RBF Networks for Learning from Partially Labeled Data. Department of Informatics, University of Klagenfurt. [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].

Erin J. Bredensteiner and Kristin P. Bennett. Multicategory Classification by Support Vector Machines. Department of Mathematics University of Evansville. [View Context].

Stefan Aeberhard and O. de Vel and Danny Coomans. New Fast Algorithms for Variable Selection based on Classifier Performance. James Cook University. [View Context].

Georg Thimm and Emile Fiesler. High Order and Multilayer Perceptron Initialization. [View Context].

Pramod Viswanath and M. Narasimha Murty and Shalabh Bhatnagar. A pattern synthesis technique to reduce the curse of dimensionality effect. E-mail. [View Context].

Chih-Wei Hsu and Cheng-Ru Lin. A Comparison of Methods for Multi-class Support Vector Machines. Department of Computer Science and Information Engineering National Taiwan University. [View Context].

Petri Kontkanen and Jussi Lahtinen and Petri Myllymaki and Tomi Silander and Henry Tirri. USING BAYESIAN NETWORKS FOR VISUALIZING HIGH-DIMENSIONAL DATA. Complex Systems Computation Group (CoSCo). [View Context].

Perry Moerland and E. Fiesler and I. Ubarretxena-Belandia. Incorporating LCLV Non-Linearities in Optical Multilayer Neural Networks. Preprint of an article published in Applied Optics. [View Context].

Matthias Scherf and W. Brauer. Feature Selection by Means of a Feature Weighting Approach. GSF - National Research Center for Environment and Health. [View Context].

Wl/odzisl/aw Duch. Coloring black boxes: visualization of neural network decisions. School of Computer Engineering, Nanyang Technological University. [View Context].

H. Altay Guvenir. A Classification Learning Algorithm Robust to Irrelevant Features. Bilkent University, Department of Computer Engineering and Information Science. [View Context].

Christian Borgelt and Rudolf Kruse. Speeding Up Fuzzy Clustering with Neural Network Techniques. Research Group Neural Networks and Fuzzy Systems Dept. of Knowledge Processing and Language Engineering, School of Computer Science Otto-von-Guericke-University of Magdeburg. [View Context].

Denver Dash and Gregory F. Cooper. Model Averaging with Discrete Bayesian Network Classifiers. Decision Systems Laboratory Intelligent Systems Program University of Pittsburgh. [View Context].

Ping Zhong and Masao Fukushima. Second Order Cone Programming Formulations for Robust Multi-class Classification. [View Context].

Aynur Akku and H. Altay Guvenir. Weighting Features in k Nearest Neighbor Classification on Feature Projections. Department of Computer Engineering and Information Science Bilkent University. [View Context].

C. Titus Brown and Harry W. Bullen and Sean P. Kelly and Robert K. Xiao and Steven G. Satterfield and John G. Hagedorn and Judith E. Devaney. Visualization and Data Mining in an 3D Immersive Environment: Summer Project 2003. [View Context].

Stefan Aeberhard and Danny Coomans and De Vel. THE PERFORMANCE OF STATISTICAL PATTERN RECOGNITION METHODS IN HIGH DIMENSIONAL SETTINGS. James Cook University. [View Context].

Pramod Viswanath and M. Narasimha Murty and Shalabh Bhatnagar. Partition Based Pattern Synthesis Technique with Efficient Algorithms for Nearest Neighbor Classification. Department of Computer Science and Automation, Indian Institute of Science. [View Context].

Yin Zhang and W. Nick Street. Bagging with Adaptive Costs. Management Sciences Department University of Iowa Iowa City. [View Context].

Daichi Mochihashi and Gen-ichiro Kikui and Kenji Kita. Learning Nonstructural Distance Metric by Minimum Cluster Distortions. ATR Spoken Language Translation research laboratories. [View Context].