1. Title of Database: Abalone data

2. Sources:

   (a) Original owners of database:
Marine Resources Division
Marine Research Laboratories - Taroona
Department of Primary Industry and Fisheries, Tasmania
GPO Box 619F, Hobart, Tasmania 7001, Australia
(contact: Warwick Nash +61 02 277277, wnash@dpi.tas.gov.au)

   (b) Donor of database:
Sam Waugh (Sam.Waugh@cs.utas.edu.au)
Department of Computer Science, University of Tasmania
GPO Box 252C, Hobart, Tasmania 7001, Australia

   (c) Date received: December 1995

3. Past Usage:

Sam Waugh (1995) "Extending and benchmarking Cascade-Correlation", PhD thesis, Computer Science Department, University of Tasmania.

-- Test set performance (final 1044 examples, first 3133 used for training):

24.86% Cascade-Correlation (no hidden nodes)
26.25% Cascade-Correlation (5 hidden nodes)
21.5% C4.5
 0.0% Linear Discriminate Analysis
 3.57% k=5 Nearest Neighbour
(Problem encoded as a classification task)

-- Data set samples are highly overlapped. Further information is required to separate completely using affine combinations. Other restrictions to data set examined.

David Clark, Zoltan Schreter, Anthony Adams "A Quantitative Comparison of Dystal and Backpropagation", submitted to the Australian Conference on Neural Networks (ACNN'96). Data set treated as a 3-category classification problem (grouping ring classes 1-8, 9 and 10, and 11 on).

-- Test set performance (3133 training, 1044 testing as above):

64%   Backprop
55%   Dystal

-- Previous work (Waugh, 1995) on same data set:

61.40% Cascade-Correlation (no hidden nodes)
65.61% Cascade-Correlation (5 hidden nodes)
59.2% C4.5
32.57% Linear Discriminate Analysis

62.46% k=5 Nearest Neighbour

4. Relevant Information Paragraph:

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

Data comes from an original (non-machine-learning) study:

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

5. Number of Instances: 4177

6. Number of Attributes: 8

7. Attribute information:

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone

Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Statistics for numeric domains:

	Length	Diam	Height	Whole	Shucked	Viscera	Shell	Rings
Min	0.075	0.055	0.000	0.002	0.001	0.001	0.002	1
Max	0.815	0.650	1.130	2.826	1.488	0.760	1.005	29
Mean	0.524	0.408	0.140	0.829	0.359	0.181	0.239	9.934
SD	0.120	0.099	0.042	0.490	0.222	0.110	0.139	3.224
Correl	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.0

8. Missing Attribute Values: None

9. Class Distribution:

Class	Examples
1	1
2	1
3	15
4	57
5	115
6	259
7	391
8	568
9	689
10	634
11	487
12	267
13	203
14	126
15	103
16	67
17	58
18	42
19	32
20	26
21	14
22	6
23	9
24	2
25	1
26	1
27	2
29	1
Total	4177