Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur	Professeur, Facultés Universitaires N.D. de la Paix, Namur
Gilles Venturini, Rapporteur	Professeur, Université de Tours
Mohand-Saïd Hacid, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Michel Lamure, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Gilbert Ritschard, Examinateur	Professeur, Université de Genève
Nicolas Nicoloyannis, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

8 Données Utilisées pour les Expérimentations

"The usefulness of data repositories like the one at UCI is subject to extreme positions : some people use them blindly without taking into account their limitations, while others simply reject them completely without trying to make adequate use of them. (...) In this paper we argue that, in principle, data repositories can be useful for KDD (...)"

- Carlos Soares -
Is the UCI Repository useful for Data Mining?
First International Workshop on Data Mining Lessons Learned (DMLL-2002)
http://www.hpl.hp.com/personal/Tom Fawcett/DMLL-2002/

Les jeux de données utilisés dans cette dissertation proviennent tous (à l'exception de quelques jeux de données synthétiques) de la collection de l'Université de Californie à IRVINE (http://www.ics.uci.edu/#mlearn/mlrepository.html) [MM96]. Le choix d'utiliser ces jeux de données est motivé ici, non pas par une croyance absolue en une qualité et une représentativité extrême de ces jeux de données, mais par la volonté de proposer des évaluations expérimentales reproductibles par d'autres chercheurs.

Nous reprenons dans les prochaines pages les descriptions des jeux de données telles qu'elles sont données dans le repertoire de l'UCI.

8.1 Jeu de Données ADULT

This data was extracted from the census bureau database found at :
http://www.census.gov/ftp/pub/DES/www/welcome.html

Donor: Ronny Kohavi and Barry Becker, Data Mining and Visualization Silicon Graphics.
e-mail: ronnyk@sgi.com for questions.
Split into train-test using MLC++ GenCV Files (2/3, 1/3 random).
48842 instances, mix of continuous and discrete (train=32561, test=16281)
45222 if instances with unknown values are removed (train=30162, test=15060)
Duplicate or conflicting instances : 6
Class probabilities for adult.all file
Probability for the label '>50K' : 23.93
Probability for the label '<=50K' : 76.07

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions:
((AAGE>16) & (AGI>100) & (AFNLWGT>1) & (HRSWK>0))
Prediction task is to determine whether a person makes over 50K a year.

First cited in: [Koh96]

Error Accuracy reported as follows, after removal of unknowns from train/test sets:
C4.5 : 84.46+-0.30
Naive-Bayes: 83.88+-0.30
NBTree : 85.90+-0.28
Following algorithms were later run with the following error rates, all after removal of unknowns and using the original train/test split. All these numbers are straight runs using MLC++ with default values.

8.2 Jeu de Données MUSHROOMS

1. Title: Mushroom Database

2. Sources:
(a) Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf
(b) Donor: Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
(c) Date: 27 April 1987

3. Past Usage:
1. Schlimmer,J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral disseration, Department of Information and Computer Science, University of California, Irvine. — STAGGER: asymptoted to 951000 instances.

2. Iba,W., Wogulis,J., & Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann. – approximately the same results with their HILLARY algorithm

3. In the following references a set of rules (given below) were learned for this data set which may serve as a point of comparison for other researchers.

Duch W, Adamczak R, Grabczewski K (1996) Extraction of logical rules from training data using backpropagation networks, in: Proc. of the The 1st Online Workshop on Soft Computing, 19-30.Aug.1996, pp. 25-30, available on-line at: http://www.bioele.nuee.nagoya-u.ac.jp/wsc1/

Duch W, Adamczak R, Grabczewski K, Ishikawa M, Ueda H, Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches, in: Proc. of the European Symposium on Artificial Neural Networks (ESANN'97), Bruge, Belgium 16-18.4.1997, pp. xx-xx

Wlodzislaw Duch, Department of Computer Methods, Nicholas Copernicus University, 87-100 Torun, Grudziadzka 5, Poland
e-mail: duch@phys.uni.torun.pl

WWW http://www.phys.uni.torun.pl/kmk/

Date: Mon, 17 Feb 1997 13:47:40 +0100
From: Wlodzislaw Duch <duch@phys.uni.torun.pl>
Organization: Dept. of Computer Methods, UMK

I have attached a file containing logical rules for mushrooms. It should be helpful for other people since only in the last year I have seen about 10 papers analyzing this dataset and obtaining quite complex rules. We will try to contribute other results later.

With best regards, Wlodek Duch

Logical rules for the mushroom data sets
Logical rules given below seem to be the simplest possible for the mushroom dataset and therefore should be treated as benchmark results.

Disjunctive rules for poisonous mushrooms, from most general to most specific:
P1) odor=NOT(almond.OR.anise.OR.none) 120 poisonous cases missed, 98.52% accuracy
P2) spore-print-color=green 48 cases missed, 99.41% accuracy
P3) odor=none.AND.stalk-surface-below-ring=scaly.AND. (stalk-color-above-ring=NOT.brown) 8 cases missed, 99.90% accuracy
P4) habitat=leaves.AND.cap-color=white 100% accuracy
Rule P4) may also be P4') population=clustered.AND.cap-color=white

These rule involve 6 attributes (out of 22). Rules for edible mushrooms are obtained as negation of the rules given above, for example the rule:
odor=(almond.OR.anise.OR.none).AND.spore-print-color=NOT.green
gives 48 errors, or 99.41% accuracy on the whole dataset.

Several slightly more complex variations on these rules exist, involving other attributes, such as gill-size, gill-spacing, stalk-surface-above-ring, but the rules given above are the simplest we have found.

4. Relevant Information:
This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

5. Number of Instances: 8124

6. Number of Attributes: 22 (all nominally valued)

7. Attribute Information: (classes: edible=e, poisonous=p)

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises?: bruises=t, no=f
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. gill-attachment: attached=a, descending=d, free=f, notched=n
7. gill-spacing: close=c, crowded=w, distant=d
8. gill-size: broad=b, narrow=n
9. gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10. stalk-shape: enlarging=e, tapering=t
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring: ibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring: ibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type: partial=p, universal=u
17. veil-color: brown=n, orange=o, white=w, yellow=y
18. ring-number: none=n, one=o, two=t
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

8. Missing Attribute Values:
2480 of them (denoted by "?"), all for attribute #11.

9. Class Distribution:
– edible: 4208 (51.8
– poisonous: 3916 (48.2
– total: 8124 instances

8.3 Jeu de Données BREAST CANCER

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 nume-

rical 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 *Proceedings of the National Academy of Sciences*, *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 *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 attribute

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

Attribute ; Domain
————————————-
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%)

8.4 Jeu de Données CAR

1. Title: Car Evaluation Database

2. Sources: (a) Creator: Marko Bohanec
(b) Donors: Marko Bohanec (marko.bohanec@ijs.si), Blaz Zupan (blaz.zupan@ijs.si)
(c) Date: June, 1997

3. Past Usage:

The hierarchical decision model, from which this dataset is derived, was first presented in : M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.

Within machine-learning, this dataset was used for the evaluation of HINT (Hierarchy INduction Tool), which was proved to be able to completely reconstruct the original hierarchical model. This, together with a comparison with C4.5, is presented in : B. Zupan, M. Bohanec, I. Bratko, J. Demsar: Machine learning by function decomposition. ICML-97, Nashville, TN. 1997 (to appear)

4. Relevant Information Paragraph:
Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX (M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.). The model evaluates cars according to the following concept structure:

CAR car acceptability . PRICE overall price . . buying buying price . . maint price of the maintenance . TECH technical characteristics . . COMFORT comfort . . . doors number of doors . . . persons capacity in terms of persons to carry . . . lug-boot the size of luggage boot . . safety estimated safety of the car

Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT. Every concept is in the original model related to its lower level descendants by a set of examples (for these examples sets see http://www-ai.ijs.si/BlazZupan/car.html).

The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug-boot, safety.

Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.

5. Number of Instances: 1728 (instances completely cover the attribute space)

6. Number of Attributes: 6

7. Attribute Values:

# Attribute ; Domain	# Attribute ; Domain
buying ; v-high, high, med, low	persons ; 2, 4, more
maint ; v-high, high, med, low	lug-boot ; small, med, big
doors ; 2, 3, 4, 5-more	safety ; low, med, high

8. Missing Attribute Values: none

9. Class Distribution (number of instances per class)

class	N	N[%]
unacc	1210	(70.023 %)
acc	384	(22.222 %)
good	69	(3.993 %)
v-good	65	(3.762 %)

8.5 Jeu de Données : ADULT

1. Title of Database: adult

2. Sources:
(a) Original owners of database (name/phone/snail address/email address) US Census Bureau.
(b) Donor of database (name/phone/snail address/email address)
Ronny Kohavi and Barry Becker,
Data Mining and Visualization Silicon Graphics.
e-mail: ronnyk@sgi.com
(c) Date received (databases may change over time without name change!) 05/19/96

3. Past Usage:
(a) Complete reference of article where it was described/used [Koh96]
(b) Indication of what attribute(s) were being predicted Salary greater or less than 50,000.
(b) Indication of study's results (i.e. Is it a good domain to use?) Hard domain with a nice number of records. The following results obtained using MLC++ with default settings for the algorithms mentioned below.

Algorithm ; Error	Algorithm Error
1 C4.5 ; 15.54	10 HOODG ; 14.82
2 C4.5-auto ; 14.46	11 FSS Naive Bayes ; 14.05
3 C4.5 rules ; 14.94	12 IDTM (Decision table) ; 14.46
4 Voted ID3 (0.6) ; 15.64	13 Naive-Bayes ; 16.12
5 Voted ID3 (0.8) ; 16.47	14 Nearest-neighbor (1) ; 21.42
6 T2 ; 16.84	15 Nearest-neighbor (3) ; 20.35
7 1R ; 19.54	16 OC1 ; 15.04
8 NBTree ; 14.10	17 Pebls ; Crashed. Unknown why (bounds WERE increased)
9 CN2 ; 16.00	

4. Relevant Information Paragraph:
Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) & (AGI>100) & (AFNLWGT>1) & (HRSWK>0))

5. Number of Instances
48842 instances, mix of continuous and discrete (train=32561, test=16281)
45222 if instances with unknown values are removed (train=30162, test=15060)
Split into train-test using MLC++ GenCVFiles (2/3, 1/3 random).

6. Number of Attributes
6 continuous, 8 nominal attributes.

7. Attribute Information:
age: continuous. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

class: >50K, <=50K

8. Missing Attribute Values: 7% have missing values.

9. Class Distribution:

Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)

Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)

8.6 Jeu de Données Contraceptive Method Choice

1. Title: Contraceptive Method Choice

2. Sources:

(a) Origin: This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey

(b) Creator: Tjen-Sien Lim (limt@stat.wisc.edu)

(c) Donor: Tjen-Sien Lim (limt@stat.wisc.edu)

(c) Date: June 7, 1997

3. Past Usage:

Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning. Forthcoming. (ftp://ftp.stat.wisc.edu/pub/loh/treeprogs/quest1.7/mach1317.pdf or (http://www.stat.wisc.edu/ limt/mach1317.pdf)

4. Relevant Information:

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not

know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics.

5. Number of Instances: 1473

6. Number of Attributes: 10 (including the class attribute)

7. Attribute Information:

1. Wife's age (numerical)
2. Wife's education (categorical) 1=low, 2, 3, 4=high
3. Husband's education (categorical) 1=low, 2, 3, 4=high
4. Number of children ever born (numerical)
5. Wife's religion (binary) 0=Non-Islam, 1=Islam
6. Wife's now working? (binary) 0=Yes, 1=No
7. Husband's occupation (categorical) 1, 2, 3, 4
8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high
9. Media exposure (binary) 0=Good, 1=Not good
10. Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

8. Missing Attribute Values: None

8.7 Jeu de Données FLAGS

1. Title: Flag database

2. Source Information
– Creators: Collected primarily from the "Collins Gem Guide to Flags": Collins Publishers (1986).
– Donor: Richard S. Forsyth
8 Grosvenor Avenue
Mapperley Park
Nottingham NG3 5DX
0602-621676
– Date: 5/15/1990

3. Past Usage:
– None known other than what is shown in Forsyth's PC/BEAGLE User's Guide.

4. Relevant Information:
– This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colours in its flag.
– 10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.

5. Number of Instances: 194

6. Number of attributes: 30 (overall)

7. Attribute Information:

1. name Name of the country concerned
2. landmass 1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania
3. zone Geographic quadrant, based on Greenwich and the Equator 1=NE, 2=SE, 3=SW, 4=NW
4. area in thousands of square km
5. population in round millions
6. language 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
7. religion 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
8. bars Number of vertical bars in the flag
9. stripes Number of horizontal stripes in the flag
10. colours Number of different colours in the flag
11. red 0 if red absent, 1 if red present in the flag
12. green same for green
13. blue same for blue
14. gold same for gold (also yellow)
15. white same for white
16. black same for black
17. orange same for orange (also brown)
18. mainhue predominant colour in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)
19. circles Number of circles in the flag
20. crosses Number of (upright) crosses
21. saltires Number of diagonal crosses
22. quarters Number of quartered sections
23. sunstars Number of sun or star symbols
24. crescent 1 if a crescent moon symbol present, else 0
25. triangle 1 if any triangles present, 0 otherwise
26. icon 1 if an inanimate image present (e.g., a boat), otherwise 0
27. animate 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
28. text 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. topleft colour in the top-left corner (moving right to decide tie-breaks)
30. botright Colour in the bottom-left corner (moving left to decide tie-breaks)

8. Missing values: None

8.8 Jeu de Données GERMAN

1. Title: German Credit data

2. Source Information

Professor Dr. Hans Hofmann Institut f"ur Statistik und "Okonometrie Universit"at Hamburg FB Wirtschaftswissenschaften Von-Melle-Park 5 2000 Hamburg 13

3. Number of Instances: 1000

Two datasets are provided. the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "german.data".

For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables.

Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by StatLog.

6. Number of Attributes : 20 (7 numerical, 13 categorical)

7. Attribute description

Attribute 1: (qualitative) Status of existing checking account : A11 : ... < 0 DM , A12 : 0 <= ... < 200 DM , A13 : ... >= 200 DM / salary assignments for at least 1 year , A14 : no checking account
Attribute 2: (numerical) Duration in month
Attribute 3: (qualitative) Credit history : A30 : no credits taken/ all credits paid back duly , A31 : all credits at this bank paid back duly , A32 : existing credits paid back duly till now , A33 : delay in paying off in the past , A34 : critical account/ other credits existing (not at this bank)
Attribute 4: (qualitative) Purpose
A40 : car (new) : A41 : car (used) , A42 : furniture/equipment , A43 : radio/television , A44 : domestic appliances , A45 : repairs , A46 : education , A47 : (vacation - does not exist?) , A48 : retraining , A49 : business , A410 : others
Attribute 5: (numerical) Credit amount
Attribute 6: (qualitative) Savings account/bonds : A61 : ... < 100 DM , A62 : 100 <= ... < 500 DM , A63 : 500 <= ... < 1000 DM , A64 : .. >= 1000 DM , A65 : unknown/ no savings account
Attribute 7: (qualitative) Present employment since : A71 : unemployed , A72 : ... < 1 year , A73 : 1 <= ... < 4 years , A74 : 4 <= ... < 7 years , A75 : .. >= 7 years
Attribute 8: (numerical) Installment rate in percentage of disposable income
Attribute 9: (qualitative) Personal status and sex : A91 : male : divorced/separated , A92 : female : divorced/separated/married , A93 : male : single , A94 : male : married/widowed , A95 : female : single
Attribute 10: (qualitative) Other debtors / guarantors : A101 : none , A102 : co-applicant , A103 : guarantor
Attribute 11: (numerical) Present residence since
Attribute 12: (qualitative) Property : A121 : real estate , A122 : if not A121 : building society savings agreement/ life insurance , A123 : if not A121/A122 : car or other, not in attribute 6 , A124 : unknown / no property
Attribute 13: (numerical) Age in years
Attribute 14: (qualitative) Other installment plans : A141 : bank , A142 : stores , A143 : none
Attribute 15: (qualitative) Housing : A151 : rent , A152 : own , A153 : for free
Attribute 16: (numerical) Number of existing credits at this bank
Attribute 17: (qualitative) Job : A171 : unemployed/ unskilled - non-resident , A172 : unskilled - resident , A173 : skilled employee / official , , A174 : management/ self-employed/ highly qualified employee/ officer
Attribute 18: (numerical) Number of people being liable to provide maintenance for
Attribute 19: (qualitative) Telephone : A191 : none , A192 : yes, registered under the customers name
Attribute 20: (qualitative) foreign worker : A201 : yes , A202 : no

8. Cost Matrix

This dataset requires use of a cost matrix (see below)

	1	2
1	0	1
2	5	0

(1 = Good, 2 = Bad)

the rows represent the actual classification and the columns the predicted classification.

It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1).

8.9 Jeu de Données HOUSE VOTES 84

1. Title: 1984 United States Congressional Voting Records Database

2. Source Information:
(a) Source: Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985.

(b) Donor: Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)

(c) Date: 27 April 1987

3. Past Usage

- Publications

1. Schlimmer, J. C. (1987). Concept acquisition through representational adjustment. Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine, CA.

– Results: about 90%-95% accuracy appears to be STAGGER's asymptote

- Predicted attribute: party affiliation (2 classes)

4. Relevant Information:

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

5. Number of Instances: 435 (267 democrats, 168) republicans)

6. Number of Attributes: 16 + class name = 17 (all Boolean valued)

7. Attribute Information:
1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)
5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)

8. Missing Attribute Values: Denoted by "?"

NOTE: It is important to recognize that "?" in this database does not mean that the value of the attribute is unknown. It means simply, that the value is not "yea" or "nay" (see "Relevant Information" section above).

Attribute: Missing Values:

1: 0 ; 2: 0 ; 3: 12 ; 4: 48 ; 5: 11 ; 6: 11 ; 7: 15 ; 8: 11 ; 9: 14 ; 10: 15 ; 11: 22 ; 12: 7 ; 13: 21 ; 14: 31 ; 15: 25 ; 16: 17 ; 17: 28

9. Class Distribution: (2 classes)

1. 45.2 percent are democrat
2. 54.8 percent are republican

8.10 Jeu de Données IONOSPHERE

1. Title: Johns Hopkins University Ionosphere database

2. Source Information:

– Donor: Vince Sigillito (vgs@aplcen.apl.jhu.edu)
– Date: 1989
– Source: Space Physics Group
Applied Physics Laboratory
Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20723

3. Past Usage:

– Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10, 262-266.

They investigated using backprop and the perceptron training algorithm on this database. Using the first 200 instances for training, which were carefully split almost 50% positive and 50% negative, they found that a "linear" perceptron attained 90.7%, a "non-linear" perceptron attained 92%, and backprop an average of over 96% accuracy on the remaining 150 test instances, consisting of 123 "good" and only 24 "bad" instances. (There was a counting error or some mistake somewhere; there are a total of 351 rather than 350 instances in this domain.) Accuracy on "good" instances was much higher than for "bad" instances. Backprop was tested with several different numbers of hidden units (in [0,15]) and incremental results were also reported (corresponding to how well the different variants of backprop did after a periodic number of epochs).

David Aha (aha@ics.uci.edu) briefly investigated this database. He found that nearest neighbor attains an accuracy of 92.1%, that Ross Quinlan's C4 algorithm attains 94.0% (no windowing), and that IB3 (Aha & Kibler, IJCAI-1989) attained 96.7% (parameter settings: 70% and 80% for acceptance and dropping respectively).

4. Relevant Information:

This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. See the paper for more details. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some

type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers for the Goose Bay system. Instances in this databse are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal.

5. Number of Instances: 351

6. Number of Attributes: 34 plus the class attribute
– All 34 predictor attributes are continuous

7. Attribute Information:
– All 34 are continuous, as described above
– The 35th attribute is either "good" or "bad" according to the definition summarized above. This is a binary classification task.

8. Missing Values: None

8.11 Jeu de Données MONKS

1. Title: The Monk's Problems

2. Sources:
(a) Donor: Sebastian Thrun
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

E-mail: thrun@cs.cmu.edu
(b) Date: October 1992

3. Past Usage:
- See File: thrun.comparison.ps.Z
- Wnek, J., "Hypothesis-driven Constructive Induction," PhD dissertation, School of Information Technology and Engineering, Reports of Machine Learning and Inference Laboratory, MLI 93-2, Center for Artificial Intelligence, George Mason University, March 1993.
- Wnek, J. and Michalski, R.S., "Comparing Symbolic and Subsymbolic Learning: Three Studies," in Machine Learning: A Multistrategy Approach, Vol. 4., R.S. Michalski and G. Tecuci (Eds.), Morgan Kaufmann, San Mateo, CA, 1993.

4. Relevant Information:
The MONK's problem were the basis of a first international comparison of learning algorithms. The result of this comparison is summarized in "The MONK's Problems - A Performance Comparison of Different Learning algorithms" by S.B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S.E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R.S. Michalski,

T. Mitchell, P. Pachowicz, Y. Reich H. Vafaie, W. Van de Welde, W. Wenzel, J. Wnek, and J. Zhang has been published as Technical Report CS-CMU-91-197, Carnegie Mellon University in Dec. 1991.

One significant characteristic of this comparison is that it was performed by a collection of researchers, each of whom was an advocate of the technique they tested (often they were the creators of the various methods). In this sense, the results are less biased than in comparisons performed by a single person advocating a specific learning method, and more accurately reflect the generalization behavior of the learning techniques as applied by knowledgeable users.

There are three MONK's problems. The domains for all MONK's problems are the same (described below). One of the MONK's problems has noise added. For each problem, the domain has been partitioned into a train and test set.

5. Number of Instances: 432

6. Number of Attributes: 8 (including class attribute)

7. Attribute information:

1. class: 0, 1
2. a1: 1, 2, 3
3. a2: 1, 2, 3
4. a3: 1, 2
5. a4: 1, 2, 3
6. a5: 1, 2, 3, 4
7. a6: 1, 2
8. Id: (A unique symbol for each instance)

8. Missing Attribute Values: None

9. Target Concepts associated to the MONK's problem:

MONK-1: (a1 = a2) or (a5 = 1)

MONK-2: EXACTLY TWO of a1 = 1, a2 = 1, a3 = 1, a4 = 1, a5 = 1, a6 = 1

MONK-3: (a5 = 3 and a4 = 1) or (a5 /= 4 and a2 /= 3) (5% class noise added to the training set)

8.12 Jeu de Données NURSERY

1. Title: Nursery Database

2. Sources:

(a) Creator: Vladislav Rajkovic et al. (13 experts)

(b) Donors: Marko Bohanec (marko.bohanec@ijs.si), Blaz Zupan (blaz.zupan@ijs.si)

(c) Date: June, 1997

3. Past Usage:

The hierarchical decision model, from which this dataset is derived, was first presented in

M. Olave, V. Rajkovic, M. Bohanec: An application for admission in public school systems. In (I. Th. M. Snellen and W. B. H. J. van de Donk and J.-P. Baquiast, editors) Expert Systems in Public Administration, pages 145-160. Elsevier Science Publishers (North Holland), 1989.

Within machine-learning, this dataset was used for the evaluation of HINT (Hierarchy INduction Tool), which was proved to be able to completely reconstruct the original hierarchical model. This, together with a comparison with C4.5, is presented in

B. Zupan, M. Bohanec, I. Bratko, J. Demsar: Machine learning by function decomposition. ICML-97, Nashville, TN. 1997 (to appear)

4. Relevant Information Paragraph:

Nursery Database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. It was used during several years in 1980's when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation. The final decision depended on three subproblems: occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The model was developed within expert system shell for decision making DEX (M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.).

The hierarchical model ranks nursery-school applications according to the following concept structure:

NURSERY Evaluation of applications for nursery schools
. EMPLOY Employment of parents and child's nursery
. . parents Parents' occupation
. . has-nurs Child's nursery
. STRUCT-FINAN Family structure and financial standings
. . STRUCTURE Family structure
. . . form Form of the family
. . . children Number of children
. . housing Housing conditions
. . finance Financial standing of the family
. SOC-HEALTH Social and health picture of the family
. . social Social conditions
. . health Health conditions

Input attributes are printed in lowercase. Besides the target concept (NURSERY) the model includes four intermediate concepts: EMPLOY, STRUCT-FINAN, STRUCTURE, SOC-HEALTH. Every concept is in the original model related to its lower level descendants by a set of examples (for these examples sets see http://www-ai.ijs.si/BlazZupan/nursery.html).

The Nursery Database contains examples with the structural information removed, i.e., directly relates NURSERY to the eight input attributes: parents, has-nurs, form, children, housing, finance, social, health.

Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods.

5. Number of Instances: 12960 (instances completely cover the attribute space)

6. Number of Attributes: 8

7. Attribute Values:
parents : usual, pretentious, great-pret
has-nurs : proper, less-proper, improper, critical, very-crit
form : complete, completed, incomplete, foster
children : 1, 2, 3, more
housing : convenient, less-conv, critical
finance : convenient, inconv
social : non-prob, slightly-prob, problematic
health : recommended, priority, not-recom

8. Missing Attribute Values: none

9. Class Distribution (number of instances per class)
class N N[%]
———————————
not-recom 4320 (33.333 %)
recommend 2 (0.015 %)
very-recom 328 (2.531 %)
priority 4266 (32.917 %)
spec-prior 4044 (31.204 %)

8.13 Jeu de Données PIMA

1. Title: Pima Indians Diabetes Database

2. Sources:
(a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases
(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231
(c) Date received: 9 May 1990

3. Past Usage:
1. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

4. Relevant Information:
Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

5. Number of Instances: 768

6. Number of Attributes: 8 plus class

7. For Each Attribute: (all numeric-valued)
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

8. Missing Attribute Values: None

9. Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

10. Brief statistical analysis:

Attribute:	Mean:	S. Deviation:	Attribute:	Mean:	S. Deviation:
1.	3.8	3.4	5.	79.8	115.2
2.	120.9	32.0	6.	32.0	7.9
3.	69.1	19.4	7.	0.5	0.3
4.	20.5	16.0	8.	33.2	11.8

8.14 Jeu de Données SICK

Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia 1987.

sick, negative. | classes
age: continuous.
sex: M, F. on
thyroxine: f, t.
query on thyroxine: f, t.
on antithyroid medication: f, t.
sick: f, t.
pregnant: f, t.
thyroid surgery: f, t.
I131 treatment: f, t.
query hypothyroid: f, t.
query hyperthyroid: f, t.
lithium: f, t.
goitre: f, t.
tumor: f, t.
hypopituitary: f, t.
psych: f, t.
TSH measured: f, t.
TSH: continuous.
T3 measured: f, t.
T3: continuous.
TT4 measured: f, t.
TT4: continuous.
T4U measured: f, t.
T4U: continuous.
FTI measured: f, t.
FTI: continuous.
TBG measured: f, t.
TBG: continuous.
referral source: WEST, STMW, SVHC, SVI, SVHD, other.

8.15 Jeu de Données SMALL SOYBEAN DISEASES

1. Title: Small Soybean Database

2. Sources:
(a) Michalski,R.S. Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss", International Journal of Policy Analysis and Information Systems, 1980, 4(2), 125-161.
(b) Donor: Doug Fisher (dfisher%vuse@uunet.uucp)
(c) Date: 1987

3. Past Usage:
1. R.S. Michalski and R.L. Chilausky "Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis", International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2, 1980.
2. Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning (pp. 121-134). Ann Arbor, Michigan: Morgan Kaufmann. – IWN recorded a 97.1% classification accuracy – 290 training and 340 test instances
3. Fisher,D.H. & Schlimmer,J.C. (1988). Concept Simplification and Predictive Accuracy. Proceedings of the Fifth International Conference on Machine Learning (pp. 22-28). Ann Arbor, Michigan: Morgan Kaufmann. – Notes why this database is highly predictable

4. Relevant Information Paragraph:
A small subset of the original soybean database. See the reference for Fisher and Schlimmer in soybean-large.names for more information.
Steven Souders wrote:
> Figure 15 in the Michalski and Stepp paper (PAMI-82) says that the > discriminant values for the attribute CONDITION OF FRUIT PODS for the > classes Rhizoctonia Root Rot and Phytophthora Rot are "few or none" > and "irrelevant" respectively. However, in the SOYBEAN-SMALL dataset > I got from UCI, the value for this attribute is "dna" (does not apply) > for both classes. I show the actual data below for cases D3 > (Rhizoctonia Root Rot) and D4 (Phytophthora Rot). According to the > attribute names given in soybean-large.names, FRUIT-PODS is attribute > #28. If you look at column 28 in the data below (marked with arrows) > you'll notice that all cases of D3 and D4 have the same value. Thus, > the SOYBEAN-SMALL dataset from UCI could NOT have produced the results > in the Michalski and Stepp paper.
I do not have that paper, but have found what is probably a later variation of that figure in Stepp's dissertation, which lists the value "normal" for the first 2 classes and "irrelevant" for the latter 2 classes. I believe that "irrelevant" is used here as a synonym for "not-applicable", "dna", and "does-not-apply". I believe that there is a mis-print in the figure he read in their PAMI-83 article.
I have checked over each attribute value in this database. It corresponds exactly with the copies listed in both Stepp's and Fisher's dissertations.

5. Number of Instances: 47

6. Number of Attributes: 35 (all have been nominalized)
– All attributes here appear with numeric values

7. Attribute Information:
Classes : diaporthe-stem-canker (D1), charcoal-rot(D2), rhizoctonia-root-rot (D3), phytophthora-rot (D4)
1. date: april,may,june,july,august,september,october,?.
2. plant-stand: normal,lt-normal,?.
3. precip: lt-norm,norm,gt-norm,?.
4. temp: lt-norm,norm,gt-norm,?.
5. hail: yes,no,?.
6. crop-hist: diff-lst-year,same-lst-yr,same-lst-two-yrs, same-lst-sev-yrs,?.
7. area-damaged: scattered,low-areas,upper-areas,whole-field,?.
8. severity: minor,pot-severe,severe,?.
9. seed-tmt: none,fungicide,other,?.
10. germination: 90-100
11. plant-growth: norm,abnorm,?.
12. leaves: norm,abnorm.
13. leafspots-halo: absent,yellow-halos,no-yellow-halos,?.
14. leafspots-marg: w-s-marg,no-w-s-marg,dna,?.
15. leafspot-size: lt-1/8,gt-1/8,dna,?.
16. leaf-shread: absent,present,?.
17. leaf-malf: absent,present,?.
18. leaf-mild: absent,upper-surf,lower-surf,?.
19. stem: norm,abnorm,?.
20. lodging: yes,no,?.
21. stem-cankers: absent,below-soil,above-soil,above-sec-nde,?.
22. canker-lesion: dna,brown,dk-brown-blk,tan,?.
23. fruiting-bodies: absent,present,?.
24. external decay: absent,firm-and-dry,watery,?.
25. mycelium: absent,present,?.
26. int-discolor: none,brown,black,?.
27. sclerotia: absent,present,?.
28. fruit-pods: norm,diseased,few-present,dna,?.
29. fruit spots: absent,colored,brown-w/blk-specks,distort,dna,?.
30. seed: norm,abnorm,?.
31. mold-growth: absent,present,?.
32. seed-discolor: absent,present,?.
33. seed-size: norm,lt-norm,?.
34. shriveling: absent,present,?.
35. roots: norm,rotted,galls-cysts,?.

8. Number of Missing Attribute Values: 0

9. Class Distribution:
1. D1: 10
2. D2: 10
3. D3: 10
4. D4: 17

8.16 Jeu de Données VEHICLE

This dataset comes from the Turing Institute, Glasgow, Scotland. If you use this dataset in any publication you must acknowledge this source.

NAME : vehicle silhouettes

PURPOSE :
to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different

angles.

PROBLEM TYPE : classification

SOURCE
Drs.Pete Mowforth and Barry Shepherd
Turing Institute
George House
36 North Hanover St.
Glasgow
G1 2AD

CONTACT
Alistair Sutherland
Statistics Dept.
Strathclyde University
Livingstone Tower
26 Richmond St.
GLASGOW G1 1XH
Great Britain
Tel: 041 552 4400 x3033
Fax: 041 552 4711
e-mail: alistair@uk.ac.strathclyde.stams

HISTORY :
This data was originally gathered at the TI in 1986-87 by JP Siebert. It was partially financed by Barr and Stroud Ltd. The original purpose was to find a method of distinguishing 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects. Measures of shape features extracted from example silhouettes of objects to be discriminated were used to generate a classification rule tree by means of computer induction. This object recognition strategy was successfully used to discriminate between silhouettes of model cars, vans and buses viewed from constrained elevation but all angles of rotation. The rule tree classification performance compared favourably to MDC (Minimum Distance Classifier) and k-NN (k-Nearest Neigh- bour) statistical classifiers in terms of both error rate and computational efficiency. An investigation of these rule trees generated by example indicated that the tree structure was heavily influenced by the orientation of the objects, and grouped similar object views into single decisions.

DESCRIPTION :
The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS, which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness. Four "Corgie" model vehicles were used for the experiment: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars. The images were acquired by a camera looking downwards at the model vehicle from a fixed angle of

elevation (34.2 degrees to the horizontal). The vehicles were placed on a diffuse backlit surface (lightbox). The vehicles were painted matte black to minimise highlights. The images were captured using a CRS4000 framestore connected to a vax 750. All images were captured with a spatial resolution of 128x128 pixels quantised to 64 greylevels. These images were thresholded to produce binary vehicle silhouettes, negated (to comply with the processing requirements of BINATTS) and thereafter subjected to shrink-expand-expand-shrink HIPS modules to remove "salt and pepper" image noise. The vehicles were rotated and their angle of orientation was measured using a radial graticule beneath the vehicle. 0 and 180 degrees corresponded to "head on" and "rear" views respectively while 90 and 270 corresponded to profiles in opposite directions. Two sets of 60 images, each set covering a full 360 degree rotation, were captured for each vehicle. The vehicle was rotated by a fixed angle between images. These datasets are known as e2 and e3 respectively. A further two sets of images, e4 and e5, were captured with the camera at elevations of 37.5 degs and 30.8 degs respectively. These sets also contain 60 images per vehicle apart from e4.van which contains only 46 owing to the difficulty of containing the van in the image at some orientations.

ATTRIBUTES :

COMPACTNESS (average perim)**2/area

CIRCULARITY (average radius)**2/area

DISTANCE CIRCULARITY area/(av.distance from border)**2

RADIUS RATIO (max.rad-min.rad)/av.radius

PR.AXIS ASPECT RATIO (minor axis)/(major axis)

MAX.LENGTH ASPECT RATIO (length perp. max length)/(max length)

SCATTER RATIO (inertia about minor axis)/(inertia about major axis)

ELONGATEDNESS area/(shrink width)**2

PR.AXIS RECTANGULARITY area/(pr.axis length*pr.axis width)

MAX.LENGTH RECTANGULARITY area/(max.length*length perp. to this)

SCALED VARIANCE (2nd order moment about minor axis)/area ALONG MAJOR AXIS

SCALED VARIANCE (2nd order moment about major axis)/area ALONG MINOR AXIS

SCALED RADIUS OF GYRATION (mavar+mivar)/area

SKEWNESS ABOUT (3rd order moment about major axis)/sigma|min**3 MAJOR AXIS

SKEWNESS ABOUT (3rd order moment about minor axis)/sigma|maj**3 MINOR AXIS

KURTOSIS ABOUT (4th order moment about major axis)/sigma|min**4 MINOR AXIS

KURTOSIS ABOUT (4th order moment about minor axis)/sigma|maj**4 MAJOR AXIS

HOLLOWS RATIO (area of hollows)/(area of bounding polygon)

Where sigma|maj**2 is the variance along the major axis and sigma|min**2 is the variance along the minor axis, and area of hollows= area of bounding polyarea of object

The area of the bounding polygon is found as a side result of the computation to find the maximum length. Each individual length computation yields a pair of calipers to the object orientated at every 5 degrees. The object is propagated into an image containing the union of these calipers to obtain an image of the bounding polygon.

NUMBER OF CLASSES : 4 OPEL, SAAB, BUS, VAN

NUMBER OF EXAMPLES :

Total no. = 946
No. in each class
opel 240 saab 240 bus 240 van 226
100 examples are being kept by Strathclyde for validation. So StatLog partners will receive 846 examples.

NUMBER OF ATTRIBUTES : No. of atts. = 18

BIBLIOGRAPHY :
Turing Institute Research Memorandum TIRM-87-018 "Vehicle Recognition Using Rule Based Methods" by Siebert,JP (March 1987)

8.17 Jeu de Données WINE

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

8.18 Jeu de Données SPAM

SPAM E-MAIL DATABASE ATTRIBUTES (in .names format)

48 continuous real [0,100] attributes of type word-freq-WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char-freq-CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital-run-length-average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital-run-length-longest = length of longest uninterrupted sequence of capital letters |

1 continuous integer [1,...] attribute of type capital-run-length-total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1 nominal 0,1 class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

For more information, see file 'spambase.DOCUMENTATION' at the UCI Machine Learning Repository: http://www.ics.uci.edu/ mlearn/MLRepository.html

classes 1, 0. (spam, non-spam)
word-freq-make: continuous.
word-freq-address: continuous.
word-freq-all: continuous.
word-freq-3d: continuous.
word-freq-our: continuous.
word-freq-over: continuous.
word-freq-remove: continuous.
word-freq-internet: continuous.
word-freq-order: continuous.
word-freq-mail: continuous.
word-freq-receive: continuous.
word-freq-will: continuous.
word-freq-people: continuous.
word-freq-report: continuous.
word-freq- addresses: continuous.
word-freq-free: continuous.
word-freq-business: continuous.
word-freq-email: continuous.
word-freq- you: continuous.
word-freq-credit: continuous.
word-freq- your: continuous.
word-freq-font: continuous.
word- freq-000: continuous.
word-freq-money: continuous.
word- freq-hp: continuous.
word-freq-hpl: continuous.
word-freq- george: continuous.
word-freq-650: continuous.
word-freq-lab: continuous.
word-freq- labs: continuous.
word-freq-telnet: continuous.
word-freq-857: continuous.
word- freq-data: continuous.
word-freq-415: continuous.
word- freq-85: continuous.
word-freq-technology: continuous.
word-freq-1999: continuous.
word-freq-parts: continuous.
word-freq-pm: continuous.
word-freq-direct: continuous.
word-freq-cs: continuous.
word-freq-meeting: continuous.
word-freq-original: continuous.
word-freq-project: continuous.
word-freq-re: continuous.
word-freq-edu: continuous.
word-freq-table: continuous.
word-freq-conference: continuous.
char-freq-;: continuous.
char-freq-(: continuous.
char-freq-[: continuous.
char-freq-!: continuous.
char-freq-: continuous.
char-freq-#: continuous.
capital-run-length-average: continuous.
capital-run-length- longest: continuous.
capital-run-length-total: continuous.