Thyroid Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 10 separate databases from Garavan Institute

Data Set Characteristics:	Multivariate, Domain-Theory	**Number of Instances:**	7200	**Area:**	Life
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	21	**Date Donated**	1987-01-01
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	149883

Source:

Ross Quinlan

Data Set Information:

From Garavan Institute
Documentation: as given by Ross Quinlan
6 databases from the Garavan Institute in Sydney, Australia
Approximately the following for each database:

** 2800 training (data) instances and 972 test instances
** Plenty of missing data
** 29 or so attributes, either Boolean or continuously-valued

2 additional databases, also from Ross Quinlan, are also here

** Hypothyroid.data and sick-euthyroid.data
** Quinlan believes that these databases have been corrupted
** Their format is highly similar to the other databases

1 more database of 9172 instances that cover 20 classes, and a related domain theory
Another thyroid database from Stefan Aeberhard

** 3 classes, 215 instances, 5 attributes
** No missing values

A Thyroid database suited for training ANNs

** 3 classes
** 3772 training instances, 3428 testing instances
** Includes cost data (donated by Peter Turney)

Attribute Information:

N/A

Relevant Papers:

Quinlan,J.R., Compton,P.J., Horn,K.A., & Lazurus,L. (1986). Inductive knowledge acquisition: A case study. In Proceedings of the Second Australian Conference on Applications of Expert Systems. Sydney, Australia. [Web Link]

Quinlan,J.R. (1986). Induction of decision trees. Machine Learning, 1, 81--106. [Web Link]

Papers That Cite This Data Set[1]:

Ken Tang and Ponnuthurai N. Suganthan and Xi Yao and A. Kai Qin. Linear dimensionalityreduction using relevance weighted LDA. School of Electrical and Electronic Engineering Nanyang Technological University. 2005. [View Context].

Vassilis Athitsos and Stan Sclaroff. Boosting Nearest Neighbor Classifiers for Multiclass Recognition. Boston University Computer Science Tech. Report No, 2004-006. 2004. [View Context].

Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004. [View Context].

Xiaoyong Chai and Li Deng and Qiang Yang and Charles X. Ling. Test-Cost Sensitive Naive Bayes Classification. ICDM. 2004. [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].

Lukasz A. Kurgan and Waldemar Swiercz and Krzysztof J. Cios. Semantic Mapping of XML Tags Using Inductive Machine Learning. ICMLA. 2002. [View Context].

Qiang Yang and Jing Wu. Enhancing the Effectiveness of Interactive Case-Based Reasoning with Clustering and Decision Forests. Appl. Intell, 14. 2001. [View Context].

Petri Kontkanen and Jussi Lahtinen and Petri Myllymäki and Henry Tirri. Unsupervised Bayesian visualization of high-dimensional data. KDD. 2000. [View Context].

Erin L. Allwein and Robert E. Schapire and Yoram Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. ICML. 2000. [View Context].

Andreas L. Prodromidis. On the Management of Distributed Learning Agents Ph.D. Thesis Proposal CUCS-032-97. Department of Computer Science Columbia University. 1998. [View Context].

Salvatore J. Stolfo and Andreas L. Prodromidis and Shelley Tselepis and Wenke Lee and David W. Fan and Philip K. Chan. JAM: Java Agents for Meta-Learning over Distributed Databases. KDD. 1997. [View Context].

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

Kai Ming Ting and Boon Toh Low. Model Combination in the Multiple-Data-Batches Scenario. ECML. 1997. [View Context].

Peter D. Turney. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. CoRR, csAI/9503102. 1995. [View Context].

George H. John and Ron Kohavi and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. ICML. 1994. [View Context].

Wl odzisl/aw Duch and Rudy Setiono and Jacek M. Zurada. Computational intelligence methods for rule-based data

understanding. [View Context].

Je Scott and Mahesan Niranjan and Richard W. Prager. Realisable Classifiers: Improving Operating Performance on Variable Cost Problems. Cambridge University Department of Engineering. [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].

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

Kai Ming Ting and Boon Toh Low. Theory Combination: an alternative to Data Combination. University of Waikato. [View Context].

Michael L. Raymer and William F. Punch and Erik D. Goodman and Leslie A. Kuhn and Anil K. Jain. Brief Papers. [View Context].

Andrew I. Schein and Lyle H. Ungar. A-Optimality for Active Learning of Logistic Regression Classifiers. Department of Computer and Information Science Levine Hall. [View Context].

Wl/odzisl/aw Duch and Rafal Adamczak and Krzysztof Grabczewski. Extraction of crisp logical rules from medical datasets. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Sherrie L. W and Zijian Zheng. A BENCHMARK FOR CLASSIFIER LEARNING. Basser Department of Computer Science The University of Sydney. [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].

Wl/odzisl/aw Duch and Rafal/ Adamczak Email:duchraad@phys. uni. torun. pl. Statistical methods for construction of neural networks. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Citation Request:

Please refer to the Machine Learning Repository's citation policy

---

[1] Papers were automatically harvested and associated with this data set, in collaboration with Rexa.info