Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 34(1), 2007, Pages 88–93
ISSN: 1223-6934

The choice of the best attribute selection measure in Decision Tree induction

Laviniu Aurelian Badulescu

Abstract. Data Mining is commonly defined as the computer-assisted search for interesting patterns and relations in large databases. Decision Trees are one of the most popular Data Mining models for classification and prediction. During the induction phase of the Decision Tree the attribute selection measure is determined by choosing the attribute that will best separate the remaining samples of the nodes partition into individual classes. The most time-consuming part of Decision Tree induction is obviously the choice of the best attribute selection measure. Thus, the choice of the best attribute selection measure is fundamental and our tests compare the performances of the 29 attribute selection measures.

2000 Mathematics Subject Classification. Primary 68T99; Secondary 62H30.
Key words and phrases. Data Mining, Decision Trees, attribute selection measure.

1. Introduction

Data Mining is a relatively young area of research that builds on the older disciplines of statistics, databases, artificial intelligence, machine learning and data visualization. The main ideas behind Data Mining are often completely opposite to mainstream statistics [15]. Data Mining is commonly defined as the computer-assisted search for interesting patterns and relations in large databases. The process must be automatic or semiautomatic. The data is always present in substantial quantities [30]. Data Mining is synonymous with Knowledge Discovery in Databases. The emergence of Data Mining is often explained by the ever increasing size of databases together with the availability of computing power and algorithms to analyze them. Data Mining is usually considered to be a form of secondary data analysis. This means that it is often performed on data collected and stored for a different purpose than analysis [12].

Decision Trees are one of the most popular Data Mining models. Decision Trees are able to provide a set of rules which improves the user's understanding. Decision Trees represent variables and variable values as trees, branches and leaves from which Decision Rules must be transformed [4]. A Decision Tree classifier builds a model by recursively dividing the training set into partitions so that all or most of the records in a partition have the same class label.

Most Decision Tree classifiers perform classification in two phases: *tree-induction* (*growing* or *building*) and *tree-pruning*. In the *tree-induction* phase the algorithm starts with the whole data set at the root node. The data set is partitioned according to a splitting criterion into subsets. This procedure is repeated recursively for each subset until each subset contains only members belonging to the same class or is sufficiently small. In the *tree-pruning* phase the full grown tree is cut back to prevent

Received: 5 June 2007.

over-fitting and to improve the accuracy of the tree [28]. During the induction phase the attribute selection measure (splitting criterion) is determined by choosing the attribute (A) that will best separate the remaining samples of the nodes partition into individual classes. This attribute becomes the decision attribute at the node. Using this attribute a splitting criterion for partitioning the data is defined, which is either of the form $A < v(v \in dom(A))$ for numeric attributes or $A \in V(V \subseteq dom(A))$ for categorical attributes. For selecting the best split point several measures were proposed (*e.g.* ID3 and C4.5 select the split that minimizes the information entropy of the partitions, while SLIQ and SPRINT use the *gini index*). Once an attribute is associated with a node, it needs not be considered in the node's children.

The most time-consuming part of Decision Tree induction is obviously the choice of the best attribute selection measure. For each active node the subset of data fulfilling the conjunction of the splitting conditions of the node and its predecessors has to be constructed and for each remaining attribute the possible splits have to be evaluated [3].

2. Performance tests

For the performance test, Decision Trees were induced on the 32,561 training records of the *Adult Database*. *Adult Database* [13] was donated by Ron Kohavi [17] and has 48,842 instances (train=32,561, test=16,281) and 15 attributes: *age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital loss, hours-per-week, native-country, class* (target attribute with two values "<=50K" and ">50K"). Missing values are confined to attributes *workclass, occupation* and *native-country*. There are 6 duplicates or conflicting instances. For the label ">50K" the probability is 23.93% and for the label "<=50K" it is 76.07%. Extraction was done by Barry Becker from the 1994 *Census database*. Prediction task is to determine whether a person makes over 50K a year. *Adult Database* was used in many others publications [22].

2.1. Attribute selection measure. There has been used 29 attribute selection measures on which the splitting of a node of the Decision Tree has to be realized. They are found in the literature, some of them being used in the induction of some very well known Decision Trees. Attribute selection measures [5, 6] used for induction, pruning and execution of Decision Trees are: information gain (*infgain*) [21, 10, 26], balanced information gain (*infgbal*), information gain ratio (*infgr*) [25, 26], symmetric information gain ratio 1 (*infsgr1*) [23], symmetric information gain ratio 2 (*infsgr2*) [23], quadratic information gain (*qigain*), balanced quadratic information gain (*qigbal*), quadratic information gain ratio (*qigr*), symmetric quadratic information gain ratio 1 (*qisgr1*), symmetric quadratic information gain ratio 2 (*qisgr2*), Gini index (*gini*) [8, 29], symmetric Gini index (*ginisym*) [31], modified Gini index (*ginimod*) [18], RELIEF measure (*relief*) [18, 16], sum of weighted differences (*wdiff*), χ^2 (*chi2*), normalized χ^2 (*chi2nrm*), weight of evidence (*wevid*) [19, 24], relevance (*relev*)[2], Bayesian-Dirichlet/K2 metric (*bdm*) [11, 9, 14], modified Bayesian-Dirichlet/K2 metric (*bdmod*) [11, 9, 14], reduction of description length - relative frequency (*rdlrel*), reduction of description length - absolute frequency (*rdlabs*), stochastic complexity (*stoco*) [20, 27], specificity gain (*spcgain*), balanced specificity gain (*spcgbal*), specificity gain ratio (*spcgr*), symmetric specificity gain ratio 1 (*spcsgr1*) [7] and symmetric specificity gain ratio 2 (*spcsgr2*) [7].

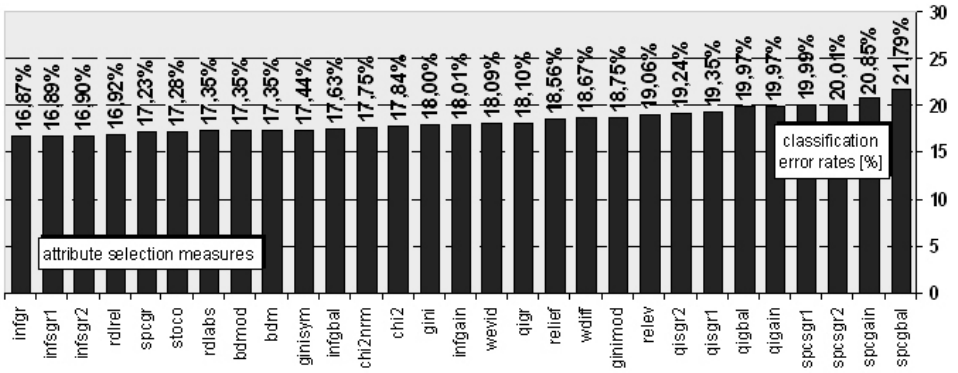

FIGURE 1. Classification error rates on the test data for unpruned Decision Trees

2.2. Decision Tree induction. Decision Trees induced at this step on the 32,561 training records of the *Adult Database* with all the 29 attribute selection measures, have been executed on the 16,281 test samples of the same database. The most important performance for the classification of the different Decision Trees, the *classification accuracy on the test data*, data completely unknown at the training of Decision Trees, has been noticed. This performance is expressed by *classification error rate* on the test data and is represented in the Figure 1 chart. In this chart, the performances are sorted in the ascending order of the classification error rates values on the test data. It can be noticed that the highest performance for the error rate on the test data is obtained by the *infgr* measure.

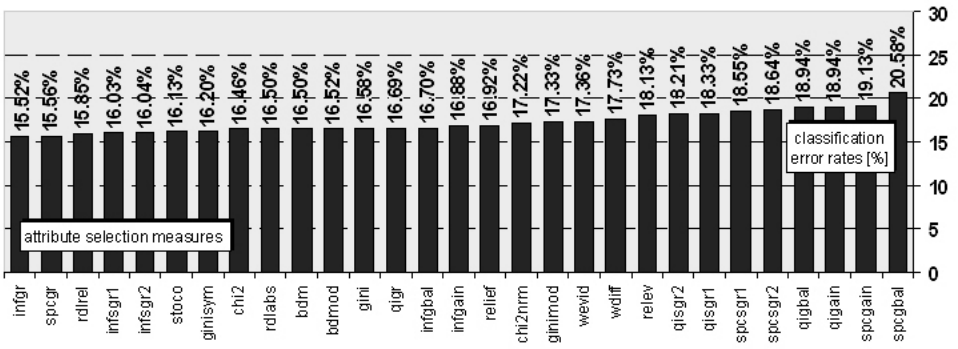

FIGURE 2. Classification error rates on the test data for pessimistic pruned Decision Trees

2.3. Decision Tree pruning with pessimistic pruning method. Decision Trees induced at the previous step was pruned by using the pessimistic pruning method. After that, Decision Trees pruned at this step was executed on the 16,281 test data of the *Adult Database*. The most important performance for the classification of the different Decision Trees, the accuracy of classification on the test data, which are

completely unknown at the Decision Trees training, is represented in the Figure 2 chart. In this chart, the performances are sorted in ascending order by values of the classification error rates. We can notice that the best performance at the classification error rate is obtained by the same *infgr* measure. For pruned Decision Trees with pessimistic pruning method the accuracy of the classification error rate is better than for unpruned Decision Trees.

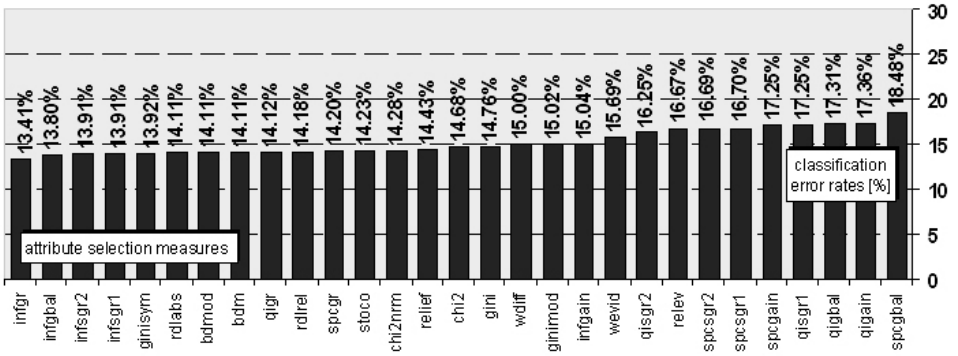

FIGURE 3. Classification error rates on the test data for confidence level pruned Decision Trees

2.4. Decision Tree pruning with confidence level pruning. Decision Trees induced at first step was pruned using confidence level pruning. Pruned Decision Trees at this step, with confidence level pruning method, was executed on the 16,281 test samples of the *Adult Database*. The accuracy of the classification on the test data is expressed in the classification error rate and is represented in the Figure 3 chart. In this chart the performances are sorted in the ascending order of the values of the classification error rates on the test data. We can notice that the best performance of the classification error rate on the test data is obtained by the same *infgr* measure. The accuracy of the classification is better than for the unpruned Decision Trees and for the Decision Trees pruned with pessimistic pruning method.

3. Conclusions

From documentation of *Adult Database*[1] we find that the following algorithms, with the classification error rates specified in square brackets: *FSS Nave Bayes* [14.05%], *NBTree* [14.10%], *C4.5-auto* [14.46%], *IDTM (Decision table)* [14.46%], *HOODG* [14.82%], *C4.5 rules* [14.94%], *OC1* [15.04%], *C4.5* [15.54%], *Voted ID3 (0.6)* [15.64%], *CN2* [16.00%], *Naive-Bayes* [16.12%], *Voted ID3 (0.8)* [16.47%], *T2* [16.84%], *1R* [19.54%], *Nearest-neighbor (3)* [20.35%], *Nearest-neighbor (1)* [21.42%], *Pebls* [Crashed], were run on *Adult* test data, all after removal of unknowns and using the original train/test split. The best performance of classification accuracy on test data is performed by *FSS Nave Bayes algorithm* with value of 14.05% for classification error rate. Our tests, using 29 attribute selection measures, find 5 attributes selection measures that outperform the best performance of the 17 algorithms presented in the documentation of *Adult Database*. Thus, for confidence level pruned Decision Trees

our tests were showed that *infgr* measure obtain an error rate of 13.41%, *infgbal* an error rate of 13.80%, *infsgr1* and *infsgr2* an error rate of 13.91%, and *ginisym* an error rate of 13.92%.

Not all the Decision Trees algorithms classify as well in any situation, but a better accuracy of the classification represents a purpose for any classifier and finally his most important performance. Thus, the choice of the best attribute selection measure is fundamental.

References

[1] *adult.names*, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/.

[2] P. W. Baim, *A method for attribute selection in inductive learning systems*, IEEE Trans. on PAMI, 10:888-896, 1988.

[3] L. A. Badulescu, *Data Mining Algorithms Based On Decision Trees*, Annals of the Oradea University. Fascicle of Management and Technological Engineering, Publishing House of Oradea University, Vol. V (XV), ISSN 1583 - 0691, pages 1621-1628, 2006.

[4] K. Blackmore, T. Bossomaier, S. Foy and D. Thomson, *Data Mining of Missing Persons Data*,S. K. Halgamuge and L. Wang (eds.), Classification and Clustering for Knowledge Discovery, Studies in Computational Intelligence, vol. 4, Springer-Verlag Berlin Heidelberg, ISBN 3-540-26073-0, page 309, 2005.

[5] C. Borgelt, *A decision tree plug-in for DataEngine*, Proc. European Congress on Intelligent Techniques and Soft Computing (EUFIT), vol. 2, pages 1299-1303, 1998.

[6] C. Borgelt, http://fuzzy.cs.uni-magdeburg.de/ borgelt/dtree.html.

[7] C. Borgelt and R. Kruse, *Evaluation Measures for Learning Probabilistic and Possibilistic Networks*, Proc. of the FUZZ-IEEE'97, vol. 2, Barcelona, Spain, pages 669-676, 1997.

[8] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Stanford University and the University of California, Berkeley, 1984.

[9] W. Buntine, *Theory Refinement on Bayesian Networks*, Proc. 7th Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufman, Los Angeles, CA, pages 52-60, 1991.

[10] C. K. Chow and C. N. Liu, *Approximating Discrete Probability Distributions with Dependence Trees*, IEEE Trans. on Information Theory, 14(3), IEEE, pages 462-467, 1968.

[11] G. F. Cooper and E. Herskovits, *A Bayesian Method for the Induction of Probabilistic Networks from Data*, Machine Learning 9, Kluwer Academic Publishers, pages 309-347, 1992.

[12] A. J. Feelders, *Data Mining in Economic Science*, J. Meij (ed.), Dealing with the data flood (STT, 65), Den Haag, the Netherlands: STT/Beweton, pages 166-175, 2002.

[13] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/.

[14] D. Heckerman, D. Geiger and D. M. Chickering, *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*, Machine Learning 20, Kluwer Academic Publishers, pages 197-243, 1995.

[15] H. C. Jessen and G. Paliouras, *Data Mining in Economics, Finance, and Marketing*, Lecture Notes in Computer Science, Vol. 2049/2001, Springer Berlin/Heidelberg, page 295, 2001.

[16] K. Kira and L. Rendell, *A practical approach to feature selection*, Proc. Intern. Conf. on Machine Learning, D. Sleeman and P. Edwards (eds.), Morgan Kaufmann, Aberdeen, pages 249-256, 1992.

[17] R. Kohavi, *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*, Proc. of the 2nd International Conf. on Knowledge Discovery and Data Mining, pages 202-207, 1996.

[18] I. Kokonenko, *Estimating Atributes: Analysis and extensions of RELIEF*, Proc. European Conf. on Machine Learning, L. De Raedt and F. Bergadano (eds.), Springer Verlag, Catania, pages 171-182, 1994.

[19] I. Kokonenko, *On Biases in Estimating Multi-Valued Attributes*, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, pages 1034-1040, 1995.

[20] R. E. Krichevsky and V. K. Trofimov, *The Performance of Universal Coding*, IEEE Trans. on Information Theory, 27(2), pages 199-207, 1983.

[21] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, Ann. Math. Statistics 22, pages 79-86, 1951.

[22] D. T. Larose, *Data Mining Methods And Models*, John Wiley and Sons, Hoboken, New Jersey, pages 18-25, 2006.

[23] R. L. de Mantaras, *A Distance-based Attribute Selection Measure for Decision Tree Induction*, Machine Learning 6, Kluwer Academic Publishers, Boston, pages 81-92, 1991.
[24] D. Michie, *Personal Models of Rationality*, Journal of Statistical Planning and Inference, Special Issue on Foundations and Philosophy of Probability and Statistics, 21, pages 381-399, 1990.
[25] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.
[26] J. R. Quinlan, *Induction of Decision Trees*, Machine Learning 1, pages 81-106, 1986.
[27] J. Rissanen, *Stochastic Complexity*, Journal of the Royal Statistical Society (Series B), vol. 49, pages 223-239, 1987.
[28] K. Uwe and S. O. Dunemann, *SQL Database Primitives for Decision Tree Classifiers*, CIKM'01 Atlanta, ACM , GA USA, 2001.
[29] L. Wehenkel, *On Uncertainty Measures Used for Decision Tree Induction*, Proc. of the International Congress on Information Processing and Management of Uncertainty in Knowledge based Systems, IPMU96, pages 413-418, 1996.
[30] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd ed., Elsevier, Morgan Kaufmann, USA, p. 5, 2005.
[31] X. Zhou and T. S. Dillon, *A statistical-heuristic Feature Selection Criterion for Decision Tree Induction*, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-13, pages 834-841, 1991.

(Laviniu Aurelian Badulescu) University of Craiova, Faculty of Automation, Computers and Electronics, Software Engineering Department,
Bvd. Decebal, Nr. 107, Craiova, Dolj, RO-200440, Romania, Tel. & Fax: 40-251438198
E-mail address: badcri@yahoo.com