

KNOWLEDGE DISCOVERY WITH DECISION RULES DERIVED FROM DECISION TREES

Lavinu Aurelian BĂDULESCU
Naiana ȚARCĂ

University of Craiova, Faculty of Automation, Computers and Electronics, Software Engineering Department, laviniu_aurelian_badulescu@yahoo.com
University of Oradea, Faculty of Economics, ntarca@uoradea.ro

Keywords: Knowledge Discovery, Decision Rules, Decision Trees

Abstract. A fundamental and well-respected technique in knowledge creation is Data Mining. Decision trees are one of the most popular Data Mining models. Any Decision Tree can be reformulated as a set of Decision Rules to make a Decision Tree more readable. The goal of our research is to analyze the performance of different Decision Trees induced with a large number of measures of split criterion in terms of training time, accuracy classification rates and the knowledge extracted in a decision rules representation.

1. INTRODUCTION

Today's economy is more and more based on knowledge and information [12]. The sharing of knowledge and creation of new knowledge leading to innovations is the key to providing future economic growth. Knowledge is now accepted as the driver of productivity and economic development, leading to a new focus on the roles of information technology and learning in economic performance. We speak about *knowledge economy* when economic growth is driven by the accumulation of knowledge, which is the basic form of capital. A knowledge driven economy is one in which the generation and exploitation of knowledge plays the predominant part in the creation of wealth [2]. Organizations trying to survive and prosper in such an economy are turning their focus to strategies, processes, tools, and technologies that can make possible the creation of knowledge. A fundamental and well-respected technique in knowledge creation is Data Mining (DM) [31].

DM is synonymous with Knowledge Discovery in Databases (KDD). DM is about extracting interesting patterns from raw data. A pattern in the data is interesting only to the extent in which it can be used in the decision-making process of the enterprise to increase utility [17]. Decision trees (DT) are one of the most popular DM models.

DT are able to provide a set of rules which improves the user's understanding. DT represent variables and variable values as trees, branches and leaves from which Decision Rules must be transformed [4]. To make a DT more readable, a path to each leaf can be transformed into an *if-then* production rule. The *if* part consists of all tests on a path, and the *then* part is a final classification. Rules in this form are called Decision Rules, and a collection of Decision Rules for all leaf nodes would classify cases exactly as the DT does [15]. Any DT can be reformulated as a set of Decision Rules [24]. The goal of our research is to analyze the performance of different DT induced with a large number of measures of split criterion in terms of training time, accuracy classification rates and the knowledge extracted in a Decision Rules representation.

Before starting the experiment, we need to specify the knowledge we want to extract, because the knowledge specificity determines what kind of mining algorithm to be chosen. In our investigation, we want to learn what kind of income classes a person should be in real world. That is a categorized problem; therefore we decide to use DT, the one of the basic techniques for data classification, to represent the knowledge that would be minded. We chose an attribute named "class $\leq 50k$, $> 50k$ " as class target attribute since we want to learn the proper classification of income for every person.

2. PERFORMANCE TESTS

For the performance tests we use software developed by C. Borgelt [6]. At first, DT was induced on the 32,561 test records of the *Adult Database*. *Adult Database* [13] was donated by Ron Kohavi [18] 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). 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 [23].

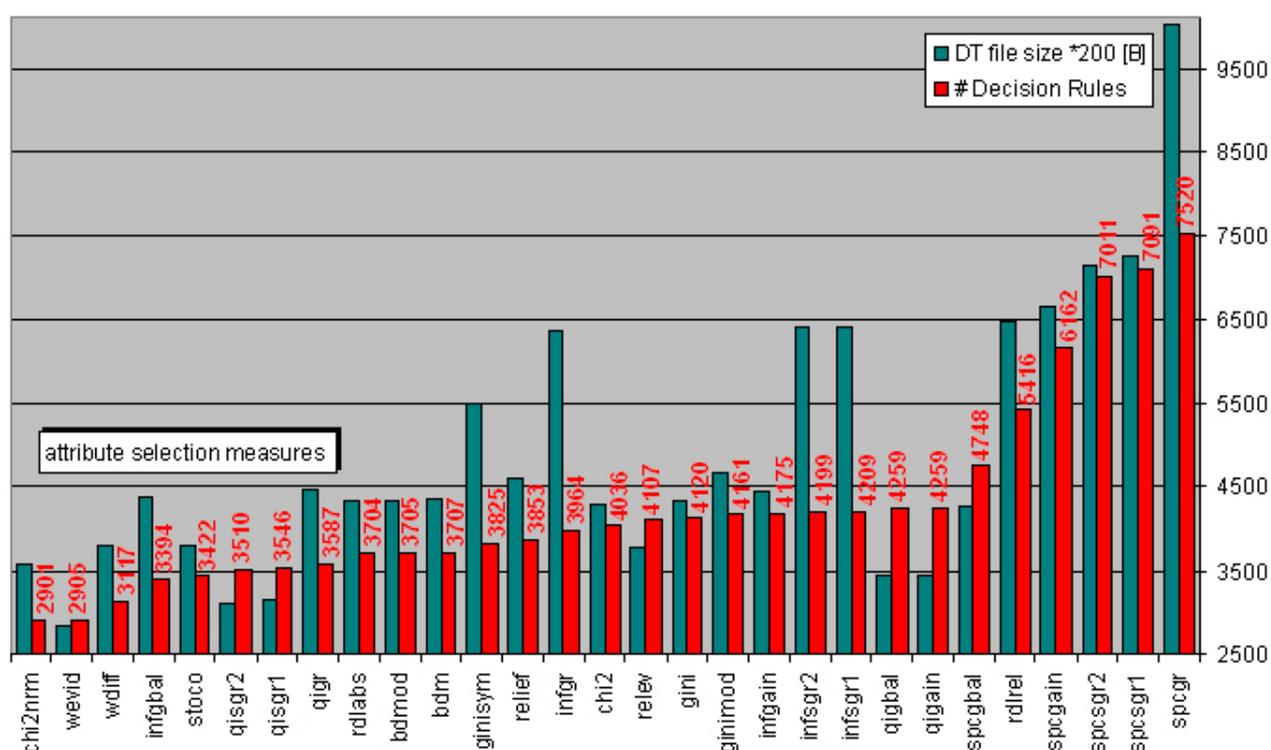


Figure 1. Extracting Decision Rules from DT induced with different attribute selection measures

There has been used 29 attribute selection measures on which the splitting of a node of the DT has to be realized. They are found in the literature, some of them being used in the induction of some very well-known DT. Attribute selection measures [5][6] used for induction, pruning and execution of DT are: information gain (*infgain*) [22][10][28], balanced information gain (*infgbal*), information gain ratio (*infr*) [27][28], symmetric information gain ratio 1 (*infsgr1*) [25], symmetric information gain ratio 2 (*infsgr2*) [25], 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][30], symmetric Gini index (*ginisym*) [32], modified Gini index (*ginimod*) [19], RELIEF measure (*relief*) [19][16], sum of weighted differences (*wdif*), χ^2 (*chi2*), normalized χ^2 (*chi2nrm*), weight of evidence (*wevid*) [20][26], relevance (*relev*) [3], 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*) [21][29], specificity gain (*spcgain*), balanced specificity gain (*spcgbal*), specificity gain ratio (*spcgr*), symmetric specificity gain ratio 1 (*spsgr1*) [7], symmetric specificity gain ratio 2 (*spsgr2*) [7].

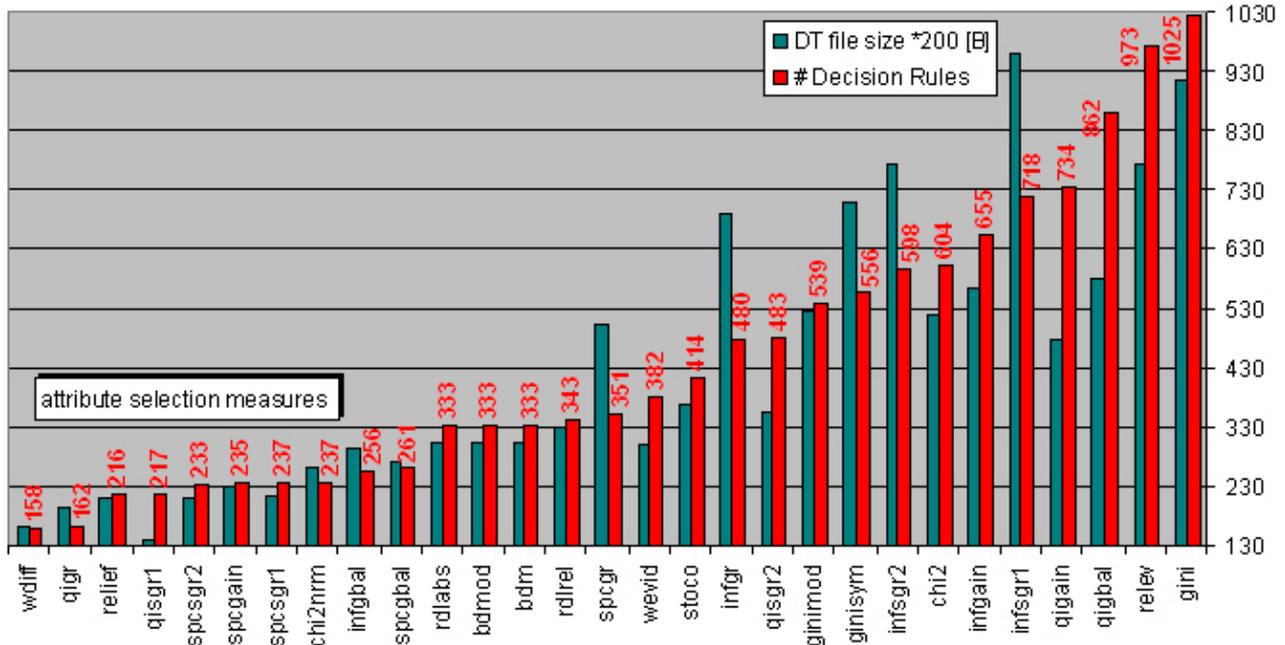


Figure 2. Extracting Decision Rules from confidence level pruned DT with different attribute selection measures

2.1. RULES EXTRACTION FROM UNPRUNED DT

Along with the performance of the Decision Rules number for every DT induced with an attribute selection measure, we have also studied the behavior of the file size performance that contains the Decision Rules. Figure 1 contains the graphical representation of the two performances for all induced DT that are based on the 29 attribute selection measures. The performances have been showed in the ascending order of the Decision Rules number.

The correlation coefficient between the number of Decision Rules for unpruned DT and the size of the file that contains these rules is high (0,830), showing a strong dependence between these performances.

2.2. RULES EXTRACTION FROM CONFIDENCE LEVEL PRUNED DT

Figure 2 contains the graphical representation of Decision Rules number and the size of file which contains these rules for all induced DT with the 29 attribute selection measures and pruned with the confidence level pruning. The performances have been showed in the ascending order of the Decision Rules number.

The correlation coefficient between the number of Decision Rules for the confidence level pruned DT and the size of the file that contains these rules, is high (0,867) showing a strong dependence between these performances, stronger that the unpruned DT.

2.3. RULES EXTRACTION FROM PESSIMISTIC PRUNED DT

Figure 3 contains the graphical representation of the Decision Rules and the size of file that contain these rules for all the DT induced on the basis of the 29 measures and pruned with the pessimistic pruning method. The performances have been sorted in an ascending order of the number of Decision Rules.

The correlation coefficient between the number of Decision Rules for the pruned DT with the pessimistic pruning method and the size of the file that contains these rules is very small (0,221) indicating a very weak dependence between these performances in contrast with unpruned DT and confidence level pruned DT to which this dependence is strong.

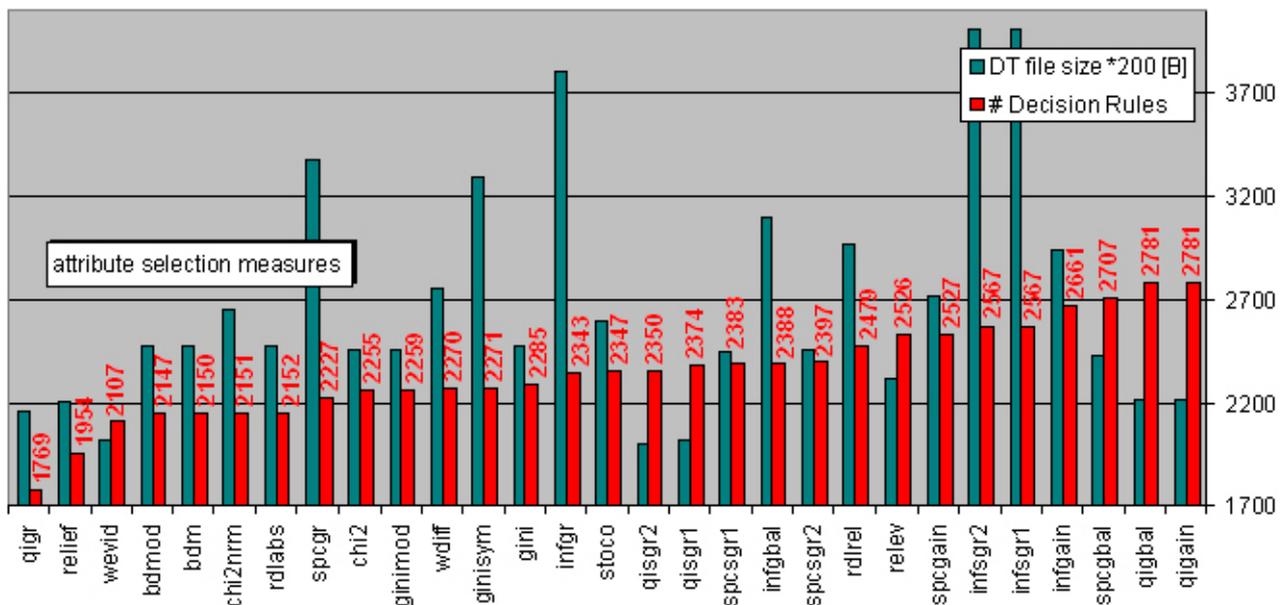


Figure 3. Extracting Decision Rules from pessimistic pruned DT with different attribute selection measures

2.4. CLASSIFICATION ERROR RATES FOR DT EXECUTION ON TEST DATA

From the three types of DT from the Decision Rules have been extracted (unpruned, confidence level pruned and pessimistic pruned) does the best classification of the DT pruned with confidence level pruning. The classification error rates have been represented in Figure 4 where the values of classification error rates for the confidence level pruning have been highlighted. The smallest classification error rate value (*i.e.* the best classification accuracy) is obtained by the DT induced with *infgr* measure, attribute selection measure used in C4.5 algorithm.

In Figure 5, we have represented the correlation coefficients between the classification error rates and the number of Decision Rules for the three types of the induced DT with 29 attribute selection measures mentioned above.

The very good correlation of all measures show that the small values for the number of Decision Rules, implies small values for the classification error rates, as it is observed at the confidence level pruned DT (see Figure 2); and the large values for the number of Decision Rules correlates with large values for classification error rate as we observe at the unpruned DT (see Figure 1).

3. CONCLUSIONS

From documentation of *Adult Database* [1] we find that the following algorithms, with the classification error rates specified in square brackets: *FSS Naïve Bayes* [14.05%],

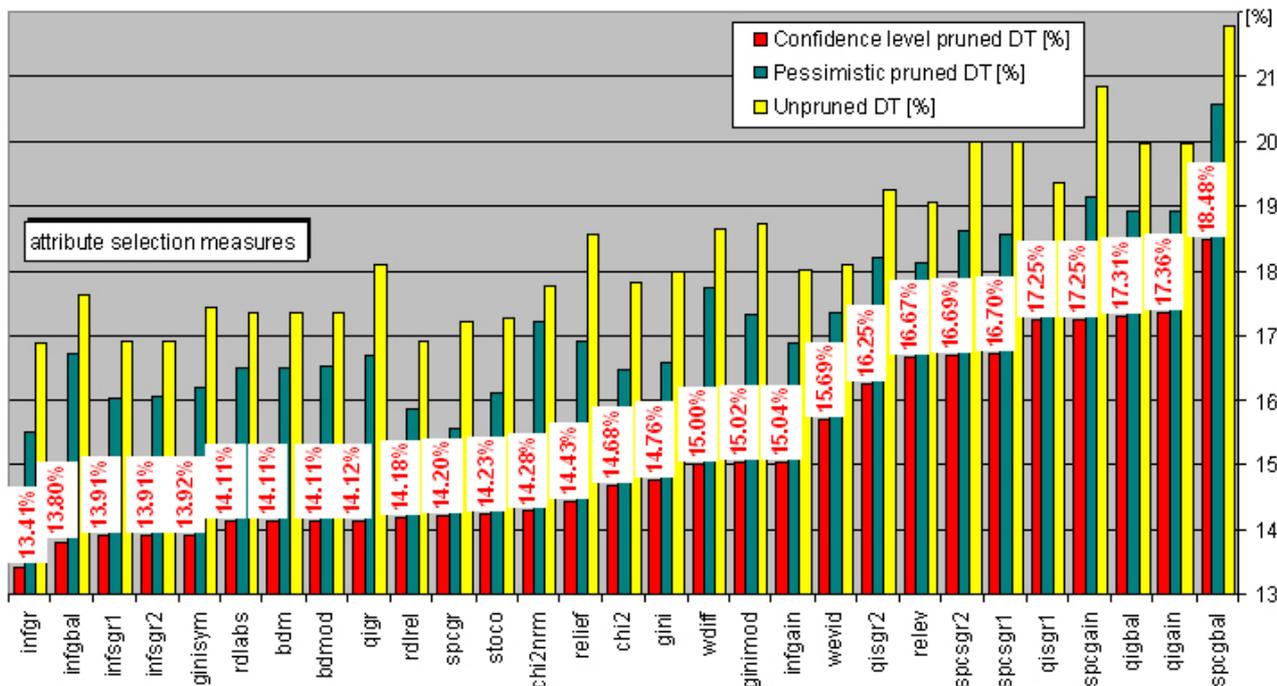


Figure 4. Classification error rates for DT execution on test data with different attribute selection measures

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 Naïve Bayes* algorithm with value of 14.05% for classification error rate.

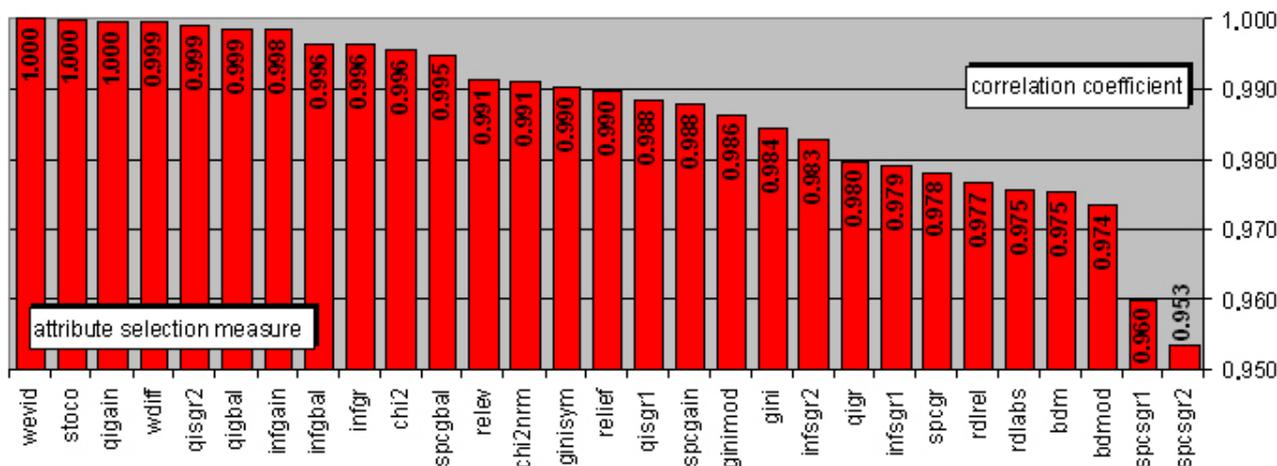


Figure 5. Correlation coefficient between classification error rates and number of Decision Rules for every attribute selection measure and every DT

Our tests find 5 attributes selection measures that outperform the best performance of the 17 algorithms presented above. Thus, for confidence level pruning DT our tests were showed that *infr* 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 DT 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. After the best DT algorithm is found, this usually means the selection of the best attribute selection measure, its representation through Decision Rules facilitate the understanding the knowledge discovery from data, knowledge so necessary in the decision-making process.

BIBLIOGRAFY

- [1] ***, *adult.names*, <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/>.
- [2] Aurum, A., *Innovative Thinking in Software Development*, in Khosrow-Pour, M., (ed.) "Encyclopedia of information science and technology", vol. III, Idea Group Inc., 2005, p. 1539.
- [3] Baim, P., W., *A method for attribute selection in inductive learning systems*, in "IEEE Trans. on PAMI", 10:888-896, 1988.
- [4] Blackmore, K., Bossomaier, T., Foy, S., Thomson, D., *Data Mining of Missing Persons Data*, in Halgamuge, S., K., Wang, L., (eds.), "Classification and Clustering for Knowledge Discovery", Studies in Computational Intelligence, vol. 4, Springer-Verlag Berlin Heidelberg, 2005, ISBN 3-540-26073-0, p. 309.
- [5] Borgelt C., *A decision tree plug-in for DataEngine*, in "Proc. European Congress on Intelligent Techniques and Soft Computing (EUFIT)", vol. 2, 1998, pages 1299-1303.
- [6] Borgelt C., <http://fuzzy.cs.uni-magdeburg.de/~borgelt/dtree.html>.
- [7] Borgelt, C., Kruse, R., *Evaluation Measures for Learning Probabilistic and Possibilistic Networks*, in "Proc. of the FUZZ-IEEE'97", vol. 2, Barcelona, Spain, 1997, pages669–676.
- [8] Breiman, L., Friedman, J., Olshen, R., Stone, C., *Classification and Regression Trees*, Stanford University and the University of California, Berkeley, 1984.
- [9] Buntine, W., *Theory Refinement on Bayesian Networks*, in "Proc. 7th Conf. on Uncertainty in Artificial Intelligence", Morgan Kaufman, Los Angeles, CA, 1991, pages 52–60.
- [10] Chow, C., K., Liu, C., N., *Approximating Discrete Probability Distributions with Dependence Trees*, in "IEEE Trans. on Information Theory", 14(3), IEEE, 1968, pages 462–467.
- [11] Cooper, G., F., Herskovits, E., *A Bayesian Method for the Induction of Probabilistic Networks from Data*, in "Machine Learning" 9, Kluwer Academic Publishers, 1992, pages 309–347.
- [12] Davenport, T., Grover, V., *Knowledge management*, in "Journal of Management Information Systems", 18(1), 2001, pages 3-4.
- [13] <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/>.
- [14] Heckerman, D., Geiger, D., Chickering, D., M., *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*, in "Machine Learning" 20, Kluwer Academic Publishers, 1995, pages197–243.
- [15] Kantardzic, M., *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, Louisville, 2003, ISBN:0471228524, Chapter 7.5.
- [16] Kira, K., Rendell, L., *A practical approach to feature selection*, in "Proc. Intern. Conf. on Machine Learning", D. Sleeman and P. Edwards (eds.), Morgan Kaufmann, Aberdeen, July 1992, pages 249-256.
- [17] Kleinberg, J., Papadimitriou, C., Raghavan, P., *A microeconomic view of data mining*, in "Knowledge Discovery and Data Mining", vol 2:4, Kluwer Academic Publishers, 1998, pages 254-260.
- [18] Kohavi, R., *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid*, in "Proc. of the 2nd International Conf. on Knowledge Discovery and Data Mining", 1996, pages 202-207.
- [19] Kokonenko, I., *Estimating Atributes: Analysis and extensions of RELIEF*, in "Proc. European Conf. on Machine Learning", L. De Raedt and F. Bergadano (eds.), Springer Verlag, Catania, April 1994, pages 171-182.
- [20] Kokonenko, I., *On Biases in Estimating Multi-Valued Attributes*, in "Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining", Montreal, 1995, pages 1034–1040.
- [21] Krichevsky, R., E., Trofimov, V., K., *The Performance of Universal Coding*, in "IEEE Trans. on Information Theory", 27(2), 1983, pages 199–207.
- [22] Kullback, S., Leibler, R., A., *On Information and Sufficiency*, in "Ann. Math. Statistics" 22, 1951, pages 79–86.
- [23] Larose, D., T., *Data Mining Methods And Models*, John Wiley & Sons, Hoboken, New Jersey, 2006, pages 18-25.

- [24] Lazar, A., *Knowledge Discovery using Heuristics*, in Khosrow-Pour, M., (ed.), "Encyclopedia of Information Science and Technology", vol. II, Idea Group Inc., USA, UK, ISBN 1-59140-553-X — ISBN 1-59140-794-X, 2005, p. 1754.
- [25] Mantaras, R., L. de, *A Distance-based Attribute Selection Measure for Decision Tree Induction*, in "Machine Learning" 6, Kluwer Academic Publishers, Boston, 1991, pages 81–92.
- [26] Michie, D, *Personal Models of Rationality*, in "Journal of Statistical Planning and Inference", Special Issue on Foundations and Philosophy of Probability and Statistics, 21, 1990, pages 381-399.
- [27] Quinlan, J., R., *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.
- [28] Quinlan, J., R., *Induction of Decision Trees*, in "Machine Learning" 1, 1986, pages 81–106.
- [29] Rissanen, J., *Stochastic Complexity*, in "Journal of the Royal Statistical Society" (Series B), vol. 49, 1987, pages 223-239.
- [30] Wehenkel, L., *On Uncertainty Measures Used for Decision Tree Induction*, in "Proc. of the International Congress on Information Processing and Management of Uncertainty in Knowledge based Systems", IPMU96, 1996, pages 413-418.
- [31] Wickramasinghe, N., *Incorporating the People Perspective into Data Mining*, in Wang, J., (ed.), "Encyclopedia of data warehousing and mining", Idea Group Inc., 2006, p. 599.
- [32] Zhou, X., Dillon, T., S., *A statistical-heuristic Feature Selection Criterion for Decision Tree Induction*, in "IEEE Trans. on Pattern Analysis and Machine Intelligence", PAMI-13, 1991, pages 834–841.