CLASSIFICATION ERROR RATES IN DECISION TREE EXECUTION

Laviniu Aurelian Badulescu

University of Craiova, Faculty of Automation, Computers and Electronics,
Software Engineering Department

Abstract: Decision Tree is a classification method used in Machine Learning and Data Mining. One major aim of a classification task is to improve its classification accuracy. In this paper, the experiments presume the induction of the different Decision Trees on four databases, using many attribute selection measures at the splitting of a Decision Tree node, the pruning of a Decision Tree using two pruning methods: confidence level pruning and pessimistic pruning method and finally, the Decision Tree execution on the test dataset to calculate the classification error rate of each Decision Tree.

Keywords: databases, machine learning, decision trees, classification, error rates.

1. INTRODUCTION

Decision tree (DT) is a classification method used in Machine Learning and Data Mining. DT helps for various decision makings. DT is grown from a training dataset having samples with several attributes. An attribute can be either a continuous attribute (*e.g.* speed, price) or a nominal attribute (*e.g.* color, country). One of the nominal attributes is designated as the class attribute; its values are called class labels. The class label indicates the class to which each sample belongs.

The result of the DT induction is symbolized as a tree which each non-leaf nodes tests an attributes and branches descending from that node specify attribute values. Leaf nodes of the tree correspond to subsets of samples with the same class label. The DT is grown by splitting the dataset at each non-leaf node according to an attribute selection measure. The primary task is to decide which of the attributes makes the best split. The best split is defined as one that does the best separating of the dataset into subsets where a single class label predominates in each subset.

The objective of classification is to use the training dataset to build a model of the class label such that it can be used to classify new samples (*e.g.* test dataset) whose class labels are unknown. The classification error rate of the training dataset should be approximately equal to the test dataset; if not, the model may be too particular for the training dataset and not general sufficient. For a classifier, classification accuracy or the capability for separating classes is a central evaluation metric of its performance.

Often, to prevent over-fitting and to improve the classification accuracy of the DT, the full grown tree is cut back in the pruning phase. Pruning phase removes subtrees that do not improve the classification accuracy. Most of pruning methods are based on minimizing a classification error rate.

One major aim of a classification task is to improve its classification accuracy. Classification accuracy estimates the degree of learning for a DT. The lower the classification error rate, the better the learning. In the same time, the enhancement of the classification accuracy improves the generalization capability of the DT. The quantitative behavior in terms of classification accuracy under different attribute selection measures and different databases can be assessed only in a large scale experiment, from which some meaningful statistics are extracted. The average

classification error rate appears to be such a meaningful statistic which also has the advantage of being simple to compute and to illustrate. The experiments we have conducted to acquire such a statistic are described next.

2. EXPERIMENTAL RESULTS

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 (Borgelt, 1998; http://fuzzy.cs.uni-magdeburg.de/~borgelt /dtree.html) used for induction, pruning and execution of DT are: information gain (*ing*) (Kullback and Leibler, 1951; Chow and Liu, 1968; Quinlan, 1986), balanced information gain (*bing*), information gain ratio (*ingr*) (Quinlan, 1993; Quinlan, 1986), symmetric information gain ratio 1 (*singr1*) (Michie, 1990), symmetric information gain ratio 2 (*singr2*) (Borgelt, 2000), quadratic information gain (*qing*) (Borgelt, 2000), balanced quadratic information gain (*bqing*), quadratic information gain ratio (*qingr*), symmetric quadratic information gain ratio 1 (*sqingr1*), symmetric quadratic information gain ratio 2 (*sqingr2*), Gini index (*gini*) (Breiman *et al*. 1984; Wehenkel, 1996), symmetric Gini index (*ginis*) (Zhou and Dillon, 1991), modified Gini index (*ginim*) (Kononenko, 1994), RELIEF measure (*relief*) (Kira and Rendell, 1992; Kononenko, 1994), sum of weighted differences (*swd*), χ^2 (*hi2*), normalized χ^2 (*hi2n*), weight of evidence (*wevd*) (Michie, 1990; Kononenko, 1995), relevance (*rlv*) (Baim, 1988), Bayesian-Dirichlet/K2 metric (*k2*) (Buntine, 1991; Cooper and Herskovits, 1992; Heckerman *et al.*, 1995), modified Bayesian-Dirichlet/K2 metric (*bd*) (Buntine, 1991; Cooper and Herskovits, 1992; Heckerman *et al.*, 1995), reduction of description length - relative frequency (*rdlrel*) (Borgelt, 2000), reduction of description length - absolute frequency (*rdlabs*) (Borgelt, 2000), stochastic complexity (*stc*) (Krichevsky and Trofimov, 1983; Rissanen, 1987), specificity gain (*sg*), balanced specificity gain (*bsg*), specificity gain ratio (*sgr*), symmetric specificity gain ratio 1 (*ssgr1*) (Borgelt and Kruse, 1997), symmetric specificity gain ratio 2 (*ssgr2*) (Borgelt and Kruse, 1997).

The experiments presume the induction of the DT on a training dataset (in fact, there were induced 29 different DT using 29 attribute selection measures at the splitting of a DT node), the pruning of a DT (the 29 DT from the previous step are pruned, using two pruning methods: confidence level pruning and pessimistic pruning method) and finally, the DT execution on the test dataset – different data of the ones used at the training of the DT - to calculate the classification error rate of each DT.

Our tests use four well-known databases from (Newman *et al.* 1998):
- *Abalone* (number of samples: 4177, 3133 training and 1044 testing; number of attributes: 8, continuous and nominal, and class attribute *rings* with values: A, B, C; missing values: none);
- *Cylinder Bands* (number of samples: 512, 412 training and 100 testing; number of attributes: 40, 20 numeric and 20 nominal, including the class attribute *band type* with values: band, no band; missing values: in 302 samples);
- *Image Segmentation* (number of samples: 6435, 4435 training and 2000 testing; number of attributes: 36, all numeric, and the class attribute with values: A, B, C, D, E and G; missing values: none) from *Statlog Project* and
- *Monk's Problem* (we use in our tests only *Monk-1* problem: number of samples: 124 training and 432 testing; number of attributes: 7, numeric, including class attribute; missing attribute values: none).

The most important performance for the classification of the different DT, the classification accuracy on the test data, data completely unknown at the training of DT, has been noticed; this performance is expressed by classification error rate on the test data.

2.1 Classification error rates for Cylinder Bands database

The classification error rates are grouped together depending on the databases, oscillating with smaller or bigger amplitudes.

Thus *Cylinder Bands database* has the biggest values for classification error rates with the biggest amplitudes between performances of different attribute selection measures.

However, the *qingr* measure reaches unexpected low values (compared with the other measures) for the classification error rate: 16% (unpruned DT) and 15% (pruned DT).

The next performance on this database is reached by the *ingr* measure (34%: the classification error rate for the unpruned DT); which is more than double that the value reached by the *qingr* measure.

If we are to compare the values of the classification error rate obtained by the *qingr* for the *Cylinder Bands database* and the three types of DT (16%, 15% and 15%), with the averages obtained by all other measures (63.38%, 68.28% and 68%), it is ascertain that the *qingr* measure's performance is about four times better.

2.2 Classification error rates for Abalone database

Abalone database has relatively big values for the classification error rate, but with small amplitudes between the different values of this performance.

2.3 Classification error rates for Image Segmentation database

Image Segmentation database has no amplitudes between the values taken by the classifications error rate for different attribute selection measures.

Thus, for unpruned DT and for pessimistic pruned DT, it is obtained the same value (16.80%) for the classification error rate, and for confidence level pruned DT the respective value is slightly smaller (15.95%).

2.4 Classification error rates for Monk's Problem database

Monk's Problem database has the best values for the

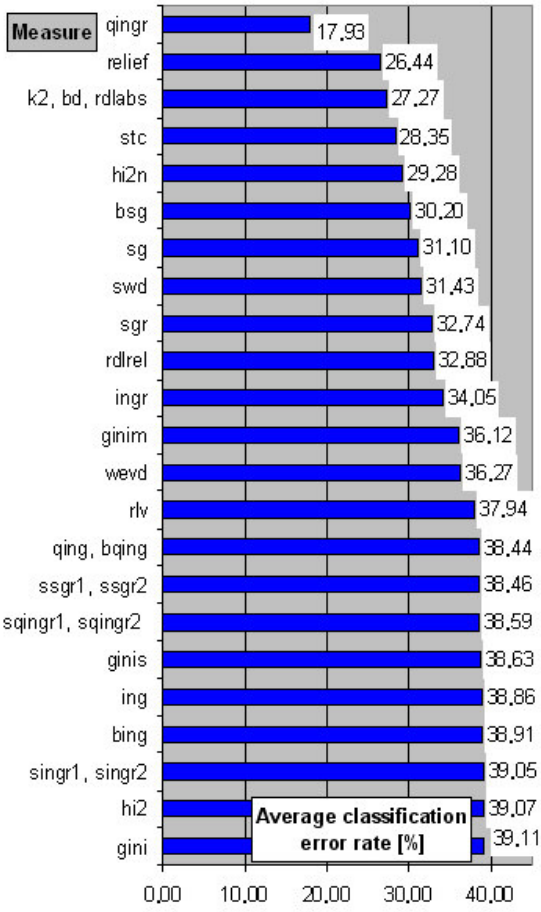

Fig. 1. The average classification error rate between the 12 types of values obtained at 4 databases and 3 types of DT

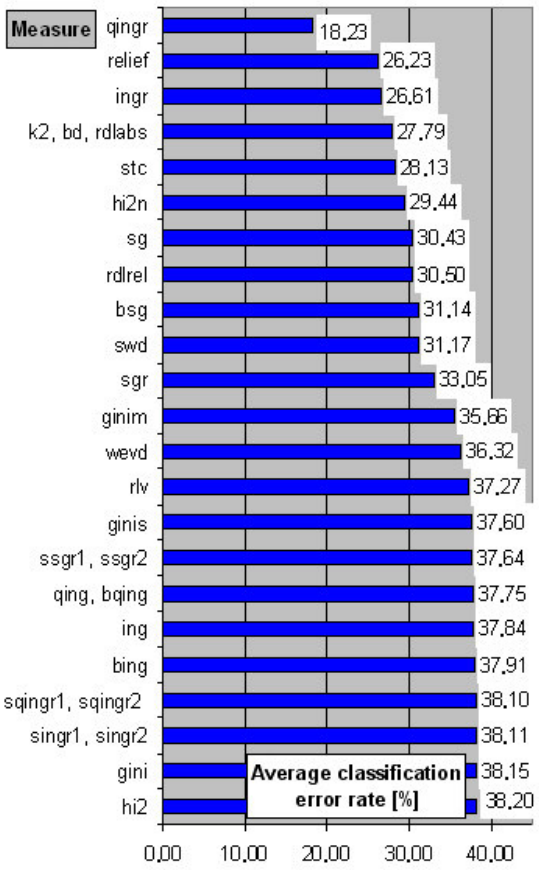

Fig. 2. The average classification error rate between the 12 types of values obtained at 4 databases only for unpruned DT

classification error rate, but with significant amplitudes between the values of the performance, though smaller than the ones from *Cylinder Bands database*.

Here the measure which systematically has the worst performance for the classification error rate on the test data is *wevd* (32.41% for unpruned DT, 27.78% for confidence level pruned DT and 36.37% for pessimistic pruned DT). To be noticed the second value which is significantly smaller then the other two.

Three measures (*qingr*, *ginim* and *relief*) make the best possible performance: 100% for classification accuracy. It is the only database, from the ones taken into consideration, where this thing happens.

Excepting these 4 measures mentioned close the other attribute selection measures slightly alternate between reasonable limits (between 8.33% and 15.74%) with big values of the classification error rate for the confidence level pruned DT.

We can say that the values of the classification error rate are almost the same at unpruned DT (average

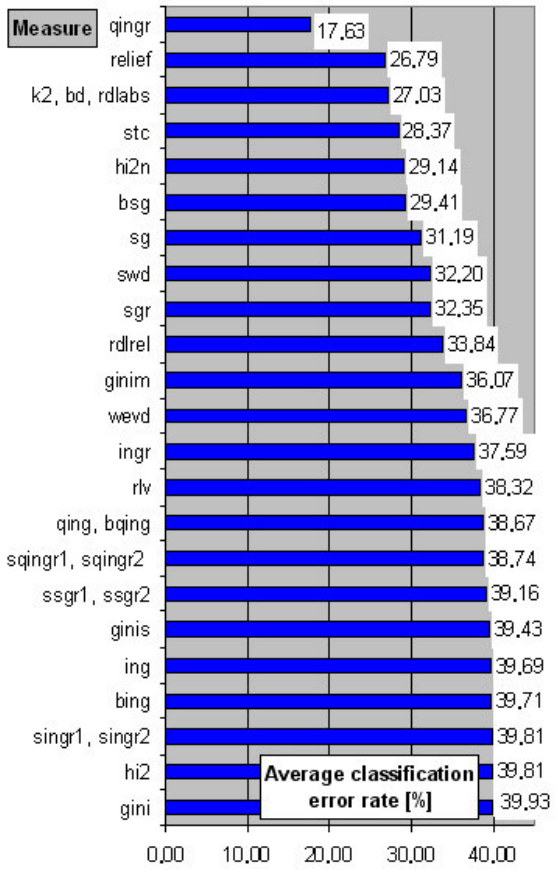

Fig. 3. The average classification error rate between the 12 types of values obtained at 4 databases only for confidence level pruned DT

10.66%) and pessimistic pruned DT (average 10.49%).

2.5. Average classification error rate

Fig. 1. presents the average accuracy of all 29 attribute selection measures for all the 4 databases taken into account and for all 3 types of DT (unpruned, confidence level pruned and pessimistic pruned).

By assuming the disadvantages which the arithmetical average presents as synthetically indicator, we can say that the *qingr* measure has a clear superior performance to any of the other 28 measures considered. Its classification error rate on the test data is with almost 10% smaller (and almost 1.5 better) than the value of the next performance, made by the *relief* measure.

The classic measures like *hi2* (CHAID algorithm) and gini (SLIQ algorithm) carry out, on the whole, the worst performances. *Ing* measure (ID3 algorithm) also carries out weak performances with 38.86%. The performance of 34.05% made by *ingr* measure (C4.5 algorithm) places it at the middle of the classification (the 11th from 23 positions).

Fig. 2. presents the average accuracy of all 29 attribute selection measures for all the 4 databases taken into account, but only for unpruned DT.

The maximum value for the average classification error rate (38.20%) is the minimum value obtained for all three types of DT (unpruned, confidence level pruned or pessimistic pruned).

Maintaining its leading position the *qingr* measure demonstrates - as we can see from the next figures (Fig. 3. and Fig. 4.) - for unpruned DT the weakest average performance for the accuracy of the classification on the test data. This performance improves as long as the pruning of the DT takes place, which is a very good thing from two points of view: we obtain a more compact DT which classifies better.

Fig. 3. presents the average accuracy of all 29 attribute selection measures for all the 4 databases taken into account, but only for confidence level pruned DT. The best performance for classification error rate for all databases used and for all types of DT tested is obtained here, for confidence level

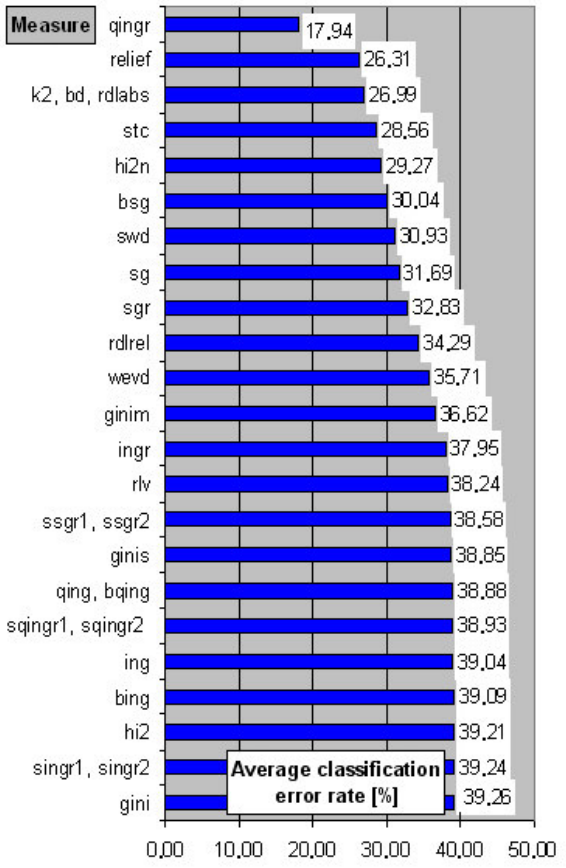

Fig 4. The average classification error rate between the 12 types of values obtained at 4 databases only for pessimistic method pruned DT

pruning (17.63%). But in the same time, the worst performance for classification error rate for all databases used and for all types of DT tested is obtained here, for confidence level pruning (39.93%).

Fig. 4. presents the average accuracy of all 29 attribute selection measures for all the 4 databases taken into account, but only for pessimistic pruned DT. For both types of pruned DT (confidence level and pessimistic) the *gini* measure occupies the last position with the worst performance for the accuracy of the classification on the test data.

3. CONCLUSIONS AND FURTHER WORK

We have investigated carefully the average classification accuracy performance of three types of DT: unpruned, confidence level pruned and pessimistic pruned. Our experiments use 29 different attribute selection measures and 4 different databases.

From all figures (Fig. 1, 2, 3 and 4) we can see that the first place is occupied by *qingr* measure and the second place is occupied by *relief* measure. Therefore we will use these two attribute selection measures for future research. For two types of pruned DT (Fig. 3 and 4) the third position is occupied by *k2*, *bd* and *rdlabs* measures. *Ingr* measure is placed on the third position for unpruned DT, but with pruning of the DT he goes down on the 13th position. The *stc* measure, with an exception (see Fig. 2, for unpruned DT, when *ingr* measure goes up on the third position and the *stc* measure occupies the 5th position) occupies the fourth position.

We must to mention a limitation of our conclusions: in evaluation by classification accuracy we have assumed equal error costs, but in the real world this is not always true. Further work is also needed to assess the performance of the 29 attribute selection measures used above on bigger datasets and with other pruning methods.

ACKNOWLEDGMENTS

We want to note the assistance we received from Newman *et al.* (1998) and Ross D. King, Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, Scotland, for the Stalog databases that are a subset of the datasets used in the European Statlog Project.

REFERENCES

Baim, P. W. (1988), A method for attribute selection in inductive learning systems, *IEEE Trans. on PAMI*, **Volume 10**, **No. 6**, pp. 888-896.

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

Borgelt, C. (2000), *Data Mining with Graphical Models*, Ph. D. Thesis, Fakultat fur Informatik der Otto-von-Guericke-Universitat Magdeburg, pp. 208, 210-211, 228.

Borgelt, C. and R. Kruse (1997), Evaluation Measures for Learning Probabilistic and Possibilistic Networks, *Proc. of the FUZZ-IEEE'97*, Barcelona, Spain, **Volume 2**, pp.669–676.

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

Buntine, W. (1991), Theory Refinement on Bayesian Networks, *Proc. 7th Conf. on Uncertainty in Artificial Intelligence (UAI 91)*, Morgan Kaufman, Los Angeles, CA, pp. 52–60.

Chow, C. K. and C. N. Liu (1968), Approximating Discrete Probability Distributions with Dependence Trees, in *IEEE Trans. on Information Theory*, **Volume 14**, **No. 3**, pages 462–467.

Cooper, G. F. and E. Herskovits (1992), A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning Journal*, Springer, **Volume 9**, **No 4**, pp. 309–347.

Heckerman, D., D. Geiger and D. M. Chickering (1995), Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning Journal* Kluwer Academic Publishers, Boston, **Volume 20**, **No. 3**, pp. 197–243.

Kira, K. and L. Rendell (1992), A practical approach to feature selection, In: *Proc. Int. Conf. on Machine Learning*, D. Sleeman and P. Edwards (Ed), pp. 249-256, Morgan Kaufmann, Aberdeen.

Kononenko, I. (1994), Estimating Atributes: Analysis and extensions of RELIEF, In: *Proc. European Conf. on Machine Learning*, L. De Raedt and F. Bergadano (Ed), pp. 171-182, Springer Verlag, Catania.

Kononenko, I. (1995), On Biases in Estimating Multi-Valued Attributes, In: *Proc. of the 14th Int. Joint Conference on Artificial Intelligence (IJCAI'95),* C. S. Mellish (Ed.), pp. 1034–1040, Morgan Kaufmann, San Mateo, CA.

Krichevsky, R. E. and V. K. Trofimov (1983), The Performance of Universal Coding, *IEEE Trans. on Information Theory*, **Volume 27**, **No 2**, pp. 199–207.

Kullback, S. and R. A. Leibler (1951), On Information and Sufficiency, *Annals of Mathematical Statistics*, **Volume 22**, **No. 1**, pages 79–86.

Michie, D. (1990), Personal Models of Rationality, *Journal of Statistical Planning and Inference, Special Issue on Foundations and Philosophy of*

Probability and Statistics, **Volume 21**, pp. 381-399.

Newman, D.J., S. Hettich, C. L. Blake and C. J. Merz (1998), *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Depart. of Information and Computer Science.

Quinlan, J. R. (1986), Induction of Decision Trees, *Machine Learning Journal*, Kluwer Academic Publishers, Boston, **Volume 1**, pp.81–106.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, Canada.

Rissanen, J. (1987), Stochastic Complexity, *Journal of the Royal Statistical Society (Series B)*, **Volume 49**, **No. 3**, pp. 223-239.

Wehenkel, L. (1996), On Uncertainty Measures Used for Decision Tree Induction, *Proc. of the Int. Congress on Information Processing and Management of Uncertainty in Knowledge based Systems (IPMU96)*, Granada, pp. 413-418.

Zhou, X. and T. S. Dillon (1991), A statistical-heuristic Feature Selection Criterion for Decision Tree Induction, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, **Volume 13, No. 8,** pp. 834–841.