Concepts for Probabilistic and Possibilistic Induction of Decision Trees on Real World Data

Christian Borgelt, Jörg Gebhardt, and Rudolf Kruse

Dept. of Mathematics and Computer Science
University of Braunschweig
D-38106 Braunschweig, Germany
e-mail: borgelt@ibr.cs.tu-bs.de

Abstract

The induction of decision trees from data is a well-known method for learning classifiers. The success of this method depends to a high degree on the measure used to select the next attribute, which, if tested, will improve the accuracy of the classification. This paper examines some possibility-based selection measures and compares them to probability- and information-based measures on real world datasets. The results show that possibility-based measures do not much worse with regard to classification accuracy, in certain cases they seem to do even slightly better.

1 Introduction

An often used method to induce decision trees from data is a greedy algorithm, which inspects the conditional distribution of the considered classes within the given dataset for each attribute. These conditional distributions are valuated with some selection measure and the attribute yielding the highest (or lowest, depending on the measure) value is chosen to partition the set of cases. Then this procedure is applied recursively on the formed subsets until all cases left in a subset are of the same class or no partition on some attribute's values leads to an improvement of the classification accuracy [12, 1]. It is obvious, that the quality of the selection measure to a high degree determines the success of this procedure.

In this paper we examine possibilistic selection measures that are based on the U-uncertainty measure of nonspecificity [6, 7]. The experimental results we obtained on real world datasets show that possibility based measures do not much worse compared to classical selection measures like information gain, information gain ratio [11] and gini index [1] as well as the g-function of Cooper and Herskovits [2], which has been used for learning Bayesian networks. In certain cases their performance is even slightly better.

2 Basic Notation

Given a set of N cases, we assume that each case is described by an instantiation of a set of attributes $\{C, A^{(1)}, \ldots, A^{(m)}\}$, where the instance of C states a class and the instances of $A^{(1)}, \ldots, A^{(m)}$ some other properties. As we will need herinafter only the class attribute C and one other attribute A, we drop the attribute index. Let the domains of C and A be defined as $\mathrm{dom}(C) = \{c_1, \ldots, c_{n_C}\}$ (a set of n_C classes) and $\mathrm{dom}(A) = \{a_1, \ldots, a_{n_A}\}$ (a set of n_A attribute values).

To state the absolute frequency of a class or an attribute value within the given set of N cases, we use the symbols $N_{i.}$ for class c_i and $N_{.j}$ for attribute value a_j. The number of cases belonging to class c_i and having attribute value a_j is denoted N_{ij} (i.e. we always use the index i to refer to a class and the index j to refer to an attribute value). The corresponding relative frequencies are written $p_{i.} = \frac{N_{i.}}{N}$, $p_{.j} = \frac{N_{.j}}{N}$ and $p_{ij} = \frac{N_{ij}}{N}$. To state the conditional relative frequency of class c_i in the subset of cases having the attribute value a_j we write $p_{i|j} = \frac{N_{ij}}{N_{.j}} = \frac{p_{ij}}{p_{.j}}$.

3 Selection Measures

3.1 Information-based Measures

One of the oldest selection measures, which was used already in ID3, is the *information gain* criterion I_{gain} [12]. It is based on Shannon entropy H [13] and defined as

$$\begin{aligned} I_{\text{gain}}(A) &= H_C - H_{C|A} \;=\; H_C + H_A - H_{CA} \\ &= -\sum_{i=1}^{n_C} p_{i.} \log_2 p_{i.} - \sum_{j=1}^{n_A} p_{.j} \log_2 p_{.j} + \sum_{i=1}^{n_C}\sum_{j=1}^{n_A} p_{ij} \log_2 p_{ij} \end{aligned}$$

Here H_C denotes the entropy of the class distribution, H_A the entropy of the attribute value distribution, and H_{CA} the entropy of the joint distribution. The idea underlying this measure can be easily understood writing $H_{C|A} = H_{CA} - H_A$ as

$$H_{C|A} \;=\; -\sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} \frac{p_{ij}}{p_{.j}} \log_2 \frac{p_{ij}}{p_{.j}} \;=\; -\sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} p_{i|j} \log_2 p_{i|j}.$$

This reveals that $H_{C|A}$ is the *expected entropy* of the class distribution with regard to attribute A, which, if compared to the entropy H_C of the unconditioned class distribution, gives the reduction of entropy (the gain of information) to be expected when the value of attribute A gets known. Hence that attribute is selected for a test which yields the highest value of I_{gain}.

It should be noted, that it would be possible to use expected entropy directly, were it impossible that values are missing from the case descriptions, because then the summand H_C would be the same for all attributes and could be discarded. But as real world datasets usually have a not neglegible number of missing values and the entropies should be calculated only on those cases, where the class as well as the attribute value is known, this summand can have different values for different attributes A, and thus it is needed to estimate the information gain correctly.

As information gain shows a strong bias in favour of multi-valued attributes [12, 8], Quinlan introduced the *information gain ratio* $I_{\text{gainratio}}$, which is defined as

$$I_{\text{gainratio}}(A) = \frac{I_{\text{gain}}(A)}{H_A} = \frac{H_C + H_A - H_{CA}}{H_A}.$$

By dividing by the entropy of the attribute value distribution the bias is strongly reduced [12, 8].

3.2 Gini Index

Another well-known selection measure for inducing decision trees is the so called gini index [1], which is defined as

$$\text{Gini}(A) \;=\; \sum_{j=1}^{n_A} p_{.j} \sum_{i=1}^{n_C} p_{i|j}^2 - \sum_{i=1}^{n_C} p_{i.}^2 \;=\; \sum_{j=1}^{n_A} \frac{1}{p_{.j}} \sum_{i=1}^{n_C} p_{ij}^2 - \sum_{i=1}^{n_C} p_{i.}^2.$$

Just as for information gain the attribute yielding the highest value is selected for a test.

3.3 *g*-function of Cooper and Herskovits

Learning decision trees has a strong resemblance to the K2 algorithm used for the induction of Bayesian networks [2], because selecting the parent nodes for some attribute in this algorithm can be seen as learning a decision tree with the special restrictions that all leaves have to lie on the same level and that all decisions on the same level of the tree have to be made on the same attribute in the same way. Hence it seems to be worth while to try the g-function defined in [2] for evaluating parent candidates as a selection measure for decision tree learning. Adapted to our notation this function reads

$$g(C, A) = c \cdot \prod_{j=1}^{n_A} \frac{(n_C - 1)!}{(N_{.j} + n_C - 1)!} \prod_{i=1}^{n_C} N_{ij}!$$

The constant c can be set to 1, as only the relation between the value of g for different attributes $A^{(1)}$ and $A^{(2)}$ matters. In order to ease the calculation of this selection measure and to make it independent of the number of cases considered, we did not use g directly, but

$$\frac{\log_2(g(C,A))}{N} = \frac{1}{N}\left(\log_2 c + \sum_{j=1}^{n_A}\left(\log_2 \frac{(n_C-1)!}{(N_{.j}+n_C-1)!} + \sum_{i=1}^{n_C}\log_2 N_{ij}!\right)\right).$$

As the g-function (for a certain value of c) estimates the probability of finding the joint distribution of C and A that is present in the dataset (assuming that the class is dependent on the value of attribute A and that all possible conditional probability distributions are equally likely), the attribute for which g (or $\log_2(g)/N$, respectively) yields the highest value is selected for a test.

3.4 Possibility-based Measures

Taking the frequency distributions as possibility distributions (an interpretation which is based on the context model of possibility theory assuming that all cases have equal weight [3, 9]), we can define possibilistic selection measures. We do so using the U-uncertainty measure of *nonspecificity* of a possibility distribution [6], which is defined as

$$\text{Nonspec}(\pi) = \int_0^{\sup(\pi)} \log_2 |[\pi]_\alpha| d\alpha$$

and can be justified as a proper generalization of Hartley information [5] to the possibilistic setting [7]. Nonspec(π) reflects the expected amount of information that has to be added in order to identify the actual value within the set $[\pi]_\alpha$ of alternatives, assuming a uniform distribution on the set $[0, \sup(\pi)]$ of possibilistic confidence levels α [4].

As the role nonspecificity plays in possibility theory is similar to that of Shannon entropy in probability theory, the idea suggests itself to construct a selection measure from it in the same way as information gain and information gain ratio are constructed from Shannon entropy.

We calculate a *specificity gain* based on the nonspecificities of the possibility distributions π_A on the set of values of attribute A, π_C on the set of classes and π_{CA} on the cartesian product of the set of values of A and the set of classes. But we have to take care how to construct these possibility distributions in order to satisfy the concepts underlying possibility theory. Just as for information gain we start from the joint distribution π_{CA} as it is induced by the given set of cases, but we calculate from it the marginal possibility distributions π_A and π_C not by summing values but by taking their maximum, i.e.

$$\begin{array}{rcll} \forall a \in A:\ \pi_A(a) &=& \max\limits_{c\in C}(\pi_{CA}(c,a)) & \qquad \text{and} \\ \forall c \in C:\ \pi_C(c) &=& \max\limits_{a\in A}(\pi_{CA}(c,a)). & \end{array}$$

In analogy to I_{gain} we then define the *specificity gain* S_{gain} as

$$S_{\text{gain}}(A) = \text{Nonspec}(\pi_C) + \text{Nonspec}(\pi_A) - \text{Nonspec}(\pi_{CA})$$

and the *specificity gain ratio* $S_{\text{gainratio}}$ as

$$S_{\text{gainratio}}(A) = \frac{S_{\text{gain}}}{\text{Nonspec}(\pi_A)} = \frac{\text{Nonspec}(\pi_C) + \text{Nonspec}(\pi_A) - \text{Nonspec}(\pi_{CA})}{\text{Nonspec}(\pi_A)}$$

The attribute which yields the highest specificity gain or specificity gain ratio is selected for a test.

Another nonspecificity measure, which we will call nonspecificity* here, we derived (with a slight modification concerning normalization) from the function m defined in [3], which was used there in an algorithm for inducing possibilistic inference networks from data. With this measure one can easily form specificity* gain and specificity* gain ratio measures in a similar way as in the preceding paragraph. A detailed definition and explanation of these measures we are forced to omit here for reasons of space.

4 Experimental Results

For our experiments on real world datasets we modified the program C4.5, release 7, by J.R. Quinlan [12], one of the most renowned decision tree learners, which is based on information gain and information gain ratio, to incorporate the selection measures presented in the preceding section. However, as C4.5 does not use information gain ratio as defined above, but includes cases with missing values in the calculation of the entropy H_A in the denominator, although they are excluded from the calculation of H_A in the numerator — a decision we do not hold to be a wise one, because it is only reasonable, if the number of missing values is small compared to the total number of cases — we reimplemented information gain ratio and did separate experiments with the original C4.5.

The experiments were conducted on some publicly accessible real world datasets from the UCI machine learning repository [10]. We present here the results on two of these: the soybean diseases dataset (table 1) and the credit card application approval dataset (table 2). The tables show for each selection measure the size of the decision tree (number of nodes including leaves), the number of errors and the error rate on training and test data before and after pruning the decision tree.

It can be seen from these results, which are quite typical, that the performance of the possibility-based measures is comparable to that of the other measures. On the test data of the soybean diseases dataset the performance of specificity gain ratio and the two specificity* measures appears to be even slightly better than that of the information-based measures. In addition, the two possibility-based measures specificity and specificity* seem to behave quite differently as on the soybean dataset specificity* yields better results, whereas on the credit card application approval the specificity measure appears to be superior. This may be due to the fact that the soybean dataset contains only nominal attributes, whereas the credit card application approval dataset also contains continuous attributes. But this may also be accidental and needs to be examined further.

Another interesting thing to observe (for which both tables are good examples) is, that decision trees grown with a possibility-based selection measure seem to suffer more from pruning. But suprisingly enough, this usually affects only the classification accuracy on the training data.

Summarizing the results, one may say that possibility-based selection measures are an interesting alternative to classical measures.

Acknowledgements

We are thankful to Prof. G. Nakhaeizadeh and his colleagues at the Daimler-Benz Research Center Ulm for supporting this project.

soybean	before pruning			after pruning		
(455/228)	size	train	test	size	train	test
default		385 (84.6)	193 (84.6)		385 (84.6)	193 (84.6)
C4.5 [release 7]	170	14 (3.1)	23 (10.1)	81	28 (6.2)	21 (9.2)
information gain	234	15 (3.3)	33 (14.5)	96	22 (4.8)	26 (11.4)
information gain ratio	164	11 (2.4)	27 (11.8)	85	18 (4.0)	22 (9.6)
gini index	278	21 (4.6)	38 (16.7)	113	38 (8.4)	33 (14.5)
$\log_2(g)/N$	326	27 (5.9)	25 (11.0)	94	53 (11.6)	31 (13.6)
specificity gain	208	11 (2.4)	29 (12.7)	84	37 (8.1)	34 (14.9)
specificity gain ratio	193	17 (3.7)	22 (9.6)	76	28 (6.2)	20 (8.8)
specificity* gain	218	6 (1.3)	15 (6.6)	81	20 (4.4)	17 (7.5)
specificity* gain ratio	151	11 (2.4)	18 (7.9)	97	22 (4.8)	19 (8.3)

Table 1: Results on the soybean diseases dataset

credit (460/230)	before pruning size	 train	 test	after pruning size	 train	 test
default		204 (44.3)	103 (44.8)		204 (44.3)	103 (44.8)
C4.5 [release 7]	89	19 (4.1)	42 (18.3)	49	28 (6.1)	36 (15.7)
information gain	106	19 (4.1)	40 (17.4)	57	28 (6.1)	39 (17.0)
information gain ratio	120	17 (3.7)	39 (17.0)	53	24 (5.2)	35 (15.2)
gini index	99	20 (4.3)	39 (17.0)	53	31 (6.7)	38 (16.5)
$\log_2(g)/N$	78	21 (4.6)	43 (18.7)	44	27 (5.9)	38 (16.5)
specificity gain	192	19 (4.1)	43 (18.7)	29	37 (8.0)	35 (15.2)
specificity gain ratio	149	23 (5.0)	41 (17.8)	31	37 (8.0)	37 (16.1)
specificity* gain	195	30 (6.5)	47 (20.4)	3	69 (15.0)	31 (13.5)
specificity* gain ratio	254	20 (4.3)	38 (16.5)	14	57 (12.4)	31 (13.5)

Table 2: Results on the credit card application approval dataset

References

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984

[2] Gregory F. Cooper and Edward Herskovits, A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning*, 9:309–347, 1992

[3] Jörg Gebhardt and Rudolf Kruse, Learning Possibilistic Networks from Data, *Proc. 5th Int. Workshop on Artificial Intelligence and Statistics*, 233–244, Fort Lauderdale, 1995

[4] Jörg Gebhardt and Rudolf Kruse, Tightest Hypertree Decompositions of Multivariate Possibility Distributions, *Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, 1996

[5] R.V.L. Hartley, Transmission of Information, *The Bell Systems Technical Journal*, 7:535–563, 1928

[6] M. Higashi and G.J. Klir, Measures of Uncertainty and Information based on Possibility Distributions, *Int. Journal of General Systems*, 9:43–58, 1982

[7] G.J. Klir and M. Mariano, On the Uniqueness of Possibility Measure of Uncertainty and Information, *Fuzzy Sets and Systems*, 24:141–160, 1987

[8] Igor Kononenko, On Biases in Estimating Multi-Valued Attributes, *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining*, 1034–1040, Montreal, 1995

[9] Rudolf Kruse, Jörg Gebhardt and Frank Klawonn, *Foundations of Fuzzy Systems*, John Wiley and Sons, Chichester, 1994

[10] P.M. Murphy and D. Aha, UCI Repository of Machine Learning Databases, FTP from ics.uci.edu in the directory pub/machine-learning-databases, 1994

[11] John Ross Quinlan, Induction of Decision Trees, *Machine Learning* 1:81–106, 1986

[12] John Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993

[13] Claude E. Shannon, The Mathematical Theory of Communication, *The Bell Systems Technical Journal*, 27:379–423, 1948