Estimating Class Probabilities in Random Forests

Henrik Boström
School of Humanities and Informatics
University of Skövde
541 28 Skövde, Sweden
henrik.bostrom@his.se

Abstract

For both single probability estimation trees (PETs) and ensembles of such trees, commonly employed class probability estimates correct the observed relative class frequencies in each leaf to avoid anomalies caused by small sample sizes. The effect of such corrections in random forests of PETs is investigated, and the use of the relative class frequency is compared to using two corrected estimates, the Laplace estimate and the m-estimate. An experiment with 34 datasets from the UCI repository shows that estimating class probabilities using relative class frequency clearly outperforms both using the Laplace estimate and the m-estimate with respect to accuracy, area under the ROC curve (AUC) and Brier score. Hence, in contrast to what is commonly employed for PETs and ensembles of PETs, these results strongly suggest that a non-corrected probability estimate should be used in random forests of PETs. The experiment further shows that learning random forests of PETs using relative class frequency significantly outperforms learning random forests of classification trees (i.e., trees for which only an unweighted vote on the most probable class is counted) with respect to both accuracy and AUC, but that the latter is clearly ahead of the former with respect to Brier score.

1. Introduction

Probability estimation trees (PETs) [15] are classification trees [6, 16] that have a class probability distribution at each leaf instead of only a single class label. Like classification trees, the PETs can be used for classifying examples, and this is simply done by assigning the most probable class according to the PET. They can also be used for ranking examples, and this is done by ordering the examples according to their likelihood of belonging to some particular class as estimated by the PET. In fact, PETs are more suited than classification trees to the latter task due to their ability to give different ranks even to examples that are assigned the same class. The former ability can be evaluated by measuring the accuracy (i.e., the percentage of correctly classified examples), which is by far the most common criterion for evaluating classifiers. There has however recently been a growing interest in also the latter ability, which can be evaluated by measuring the area under the ROC curve (AUC) [10]. The AUC can be interpreted as the probability of ranking a true positive example ahead of a false positive when ordering examples according to decreasing likelihood of being positive [3]. One reason for choosing to compare models with respect to AUC instead of accuracy is that the former is not sensitive to differences between the class distribution of the training examples and of the examples on which the model is applied [3, 9]. This means that when using AUC instead of accuracy for comparing models, one is less likely being mislead when choosing a model due to having evaluated the model on a skewed sample. Yet another measure for evaluating PETs is the mean squared error of the assigned class probabilities, a measure also known as the Brier score [7]. It can be considered to give an estimate of how reliable the assigned class probabilities are, something which can be of major importance in many applications. It should be noted that these measures are not completely correlated, and a model that is less accurate than another, may very well result in a higher AUC or Brier score (and vice versa).

In this work, we consider ensembles of PETs which, similarly to ensembles of ordinary classification trees, have been shown to consistently outperform single trees [15]. One finding that is common for both ensembles of classification trees and ensembles of PETs is that pruning has a detrimental effect [1]. This can be explained by that the ensemble in fact exploits the variance (or diversity) of the individual models, something which is reduced by pruning. One consequence of not employing pruning in a PET is that the resulting class probability distributions often have to be formed from very few training examples, sometimes even a single example. In such cases, the use of the relative class frequency as a probability estimate may lead to obvious anomalies. For example, an example that falls into a leaf node for which only a single example has been used to estimate class probabilities will obtain the highest possible probability for one of the classes, and it will be ranked ahead of any other example that falls into a leaf for which at least two differently labeled examples have been used to estimate class probabilities, independently of the number of examples and how the classes are distributed among them (e.g., an example that falls into a leaf for which there are 99 positive and 1 negative example will be ranked as less likely positive compared to an example that falls into a leaf for which there is a single positive example only). A commonly proposed remedy to this problem is to use some correction, such as the Laplace correction or the m-estimate [15, 12, 13], that pushes class probability distributions that are inferred from few observations towards what can be a priori expected. In a way, such a correction tries to reduce error due to variance by introducing a bias. However, although the use of corrected probabilities has been demonstrated to be beneficial for single PETs, as has the use of ensembles for PETs, the use of corrected probability estimates in conjunction with ensembles of PETs has, according to the best of our knowledge, not been compared to ensembles of PETs without such corrections. Still, corrected probability estimates have been widely adopted also for ensembles of PETs [15, 13, 14]. This study aims to bring some light on whether corrected probability estimates indeed are beneficial for random forests of PETs.

In the next section, we recapitulate the most commonly used probability estimates, which are to be investigated in conjunction with random forests. In Section 3, we describe the experimental setup for this study, present results from comparing the probability estimates with regard to accuracy, AUC and Brier score, and provide some explanations for the observed differences in performance. Finally, we give some concluding remarks and outline future work in Section 4.

2. Probability Estimates

In this section, we present four different ways of estimating class probabilities in an ensemble of PETs: one averaging the votes of the members of the ensemble, where each member contributes with an unweighted vote for a single class (hence acting as a classification tree), and the remaining three averaging class probability distributions estimated by the relative class frequency, the Laplace estimate and the m-estimate respectively.

2.1. Average Vote

The *average vote* defines a class probability distribution by averaging the unweighted class votes by the members of the ensemble, where each member vote for a single (most probable) class. It should be noted that the actual class probability distributions of the members are not used other than for choosing the most probable class. This means that the average vote for ensembles of PETs gives the same result as for ensembles of classification trees for which the class probability distributions have been replaced by the most probable class. The average vote (AV) can be defined as:

$$AV(\{t_1, \ldots, t_N\}, e, k) = \frac{\sum_{i=1}^{N} 1(max_{k'}\{t_i(e, k')\} = k)}{N}$$

where $t_1, \ldots, t_N$ are the members (PETs) of the ensemble, e is the example to be classified, k is a class label, and where each $t_i(e, k')$ returns the estimated probability of e belonging to class k' according to t_i. The function $1(s)$ returns 1 if s is true, and 0 otherwise.

2.2. Relative Class Frequency

The *relative class frequency* defines a class probability distribution by averaging the relative class frequencies of the members of the ensemble. The relative class frequency (RF) can be defined as:

$$RF(\{t_1, \ldots, t_N\}, e, k) = \frac{\sum_{i=1}^{N} rf(t_i, e, k)}{N}$$

where again $t_1, \ldots, t_N$ are the members of the ensemble, e is the example to be classified, k is a class label, and where $rf(t_i, e, k)$ is the relative class frequency of k in the leaf node into which e falls:

$$rf(t, e, k) = \frac{l(t, e, k)}{\sum_{j=1}^{K} l(t, e, k_j)}$$

where $l(t, e, k)$ gives the number of *estimation examples* (i.e., the set of examples that is used for estimating probabilities) belonging to class k that falls into the same leaf as example e in t. It should be noted that there are several strategies for selecting estimation examples. In the case of generating only a single tree, the estimation examples are most commonly chosen to be separate from the set of examples that is used to grow the tree. On the other hand, for bagged trees [4] and random forests [5], this set is commonly chosen to be the same as the original training set (although the alternative of using the *out-of-bag* examples has also been explored, but with less success [14]). In this work, we have adopted the common strategy of choosing the estimation examples to be identical to the original set of training examples.

2.3. Laplace estimate

The *Laplace estimate* defines a class probability distribution by averaging probability estimates that adjust the relative frequencies by adding one to the number of observed estimation examples for each class in each leaf. The Laplace estimate (LP) can be defined as:

$$LP(\{t_1, \ldots, t_N\}, e, k) = \frac{\sum_{i=1}^{N} lp(t_i, e, k)}{N}$$

where again $t_1, \ldots, t_N$ are the members of the ensemble, e is the example to be classified, k is a class label, and where $lp(t_i, e, k)$ gives the Laplace corrected probability of e belonging to class k according to t_i:

$$lp(t, e, k) = \frac{1 + l(t, e, k)}{K + \sum_{j=1}^{K} l(t, e, k_j)}$$

where $l(t, e, k)$ again gives the number of estimation examples belonging to class k that falls into the same leaf as example e in t, and where K is the number of classes.

2.4. The m-estimate

The *m-estimate* defines a class probability distribution by averaging probability estimates that adjust the relative frequencies by adding to each leaf m estimation examples distributed according to the *a priori* class probability distribution, where m is a parameter of the estimate. The m-estimate (M=m) can be defined as:

$$M_m(\{t_1, \ldots, t_N\}, e, k) = \frac{\sum_{i=1}^{N} p_m(t_i, e, k)}{N}$$

where $t_1, \ldots, t_N$ are the members of the ensemble, e is the example to be classified, k is a class label, and where $p_m(t_i, e, k)$ is the corrected probability of e belonging to class k according to t_i:

$$p_m(t, e, k) = \frac{m \times P_k + l(t, e, k)}{m + \sum_{j=1}^{K} l(t, e, k_j)}$$

where $l(t, e, k)$ again gives the number of estimation examples belonging to class k that falls into the same leaf as example e in t, and P_k is the *a priori* probability for class k. The latter is typically estimated using all available training examples, an approach which is also adopted in this study.

3. Empirical Analysis

3.1. Experimental Setting

3.1.1 Methods

The probability estimates that are to be compared for random forests [5] are: average vote (AV), relative class frequency (RF), the Laplace estimate (LP) and the m-estimate (M=m). We have explored three settings for the latter: $m = 1$, $m = 2$ and $m = K$ (no. of classes). It should be noted that for the last two settings, the m-estimate coincides with the Laplace estimate in case all classes are distributed equally, and that the second and third setting coincide for binary classification tasks. AV assumes each member votes for the most frequent class in the estimation set of the corresponding leaf (i.e., this corresponds to choosing the most probable class according to relative frequency or Laplace).

All probability estimates are used in conjunction with random forests consisting of 50 trees. Each tree is generated from a bootstrap replicate of the training set [4], and at each node in the tree generation, only a random subset of the available attributes are considered for partitioning the examples, where the size of this subset is set to be equal to the square root of the number of available attributes (as suggested in [5]). The entire set of training examples is used as estimation examples, as discussed in Section 2.2. All compared ensembles are identical except for the class probability estimates that are used when classifying novel instances.

3.1.2 Methodology and data sets

The methods are compared w.r.t. accuracy, AUC and Brier score using stratified ten-fold cross-validation on 34 data sets from the UCI Repository [2]. The names of the data sets together with the number of classes are listed in Table 2. The AUC was calculated for each method on all examples according to [11]. For data sets with more than two classes, the total AUC was calculated [10].[1]

3.1.3 Test hypotheses

There are actually a number of hypotheses to be tested. The null hypotheses can be formulated as there is no difference in predictive performance (i.e, as measured by accuracy, AUC and Brier score) between random forests of PETs using relative frequency, the Laplace estimate and the m-estimate, and random forests of classification trees.

[1] For two-class problems, the total AUC is equivalent to AUC.

3.2. Experimental Results

3.2.1 Accuracy

The number of wins and losses for each pair of methods with respect to accuracy is shown in Table 1(a), where results for which the p-value (double-sided binomial tail probability) is less than 0.05 are marked with bold-face. The actual accuracies obtained for all methods on the 34 data sets are shown in Table 2.

Random forests of PETs using relative frequency turns out to be a clear winner, with a won/loss ratio against both average vote and Laplace significantly deviating from the expected if the null hypotheses were true (the p-values for obtaining these ratios if the null hypotheses hold are 2.49×10^{-03} and 3.47×10^{-02} respectively). Relative frequency is also clearly ahead of the best setting for m (although the p-value of 0.12 is not below the significance threshold).

3.2.2 The area under the ROC curve

The number of wins and losses for each pair of methods with respect to AUC is shown in Table 1(b), with the actual AUC values shown in Table 2.

Again, random forests of PETs using relative frequency is a clear winner, although won/loss ratios are only significant compared to average vote and the two last settings for the m-estimate (the p-values for obtaining these ratios are 1.32×10^{-03} and 1.35×10^{-02} respectively). Relative frequency is also clearly ahead of the best setting for m (p-value of 0.16) and Laplace (0.30), although these are not below the chosen significance threshold.

3.2.3 Brier score

The number of wins and losses for each pair of methods regarding Brier score is shown in Table 1(c), and the actual scores in Table 3.

The use of relative frequency for random forests of PETs significantly outperforms random forests of PETs using Laplace and the m-estimate (for all three settings). The p-values for obtaining the observed won/loss ratios under the null hypotheses are 4.07×10^{-09} and 6.94×10^{-08} respectively, allowing the null hypotheses to be safely rejected. In contrast to the results for the previous performance measures, the clear winner is however random forests of classification trees - significantly outperforming both Laplace and m-estimate (the p-values are 1.95×10^{-04} when compared to Laplace and 2.94×10^{-03} when compared to the best value for the m-estimate, i.e., $m = 1$), and leaving relative frequency clearly behind (the p-value of 0.12 is however not significant).

Table 1. Wins and losses (row wins/column wins).

(a) Accuracy

	AV	RF	LP	M=1	M=2	M=K
AV	-	**5/21**	10/16	9/18	13/15	17/12
RF	**21/5**	-	**17/6**	18/9	**21/6**	**22/6**
LP	16/10	**6/17**	-	10/14	14/10	17/8
M=1	18/9	9/18	14/10	-	**17/6**	**20/5**
M=2	15/13	**6/21**	10/14	**6/17**	-	10/3
M=K	12/17	**6/22**	8/17	**5/20**	3/10	-

(b) AUC

	AV	RF	LP	M=1	M=2	M=K
AV	-	**7/26**	11/22	10/22	**10/23**	13/20
RF	**26/7**	-	20/13	21/12	**24/9**	**24/9**
LP	22/11	13/20	-	14/19	19/13	**24/6**
M=1	22/10	12/21	19/14	-	22/10	**23/10**
M=2	**23/10**	**9/24**	13/19	10/22	-	**15/3**
M=K	20/13	**9/24**	**6/24**	**10/23**	**3/15**	-

(c) Brier score

	AV	RF	LP	M=1	M=2	M=K
AV	-	22/12	**28/6**	**26/8**	**27/7**	**29/5**
RF	12/22	-	**33/1**	**32/2**	**32/2**	**32/2**
LP	**6/28**	**1/33**	-	**1/33**	14/20	20/12
M=1	**8/26**	**2/32**	**33/1**	-	**32/2**	**32/2**
M=2	**7/27**	**2/32**	20/14	**2/32**	-	**18/1**
M=K	**5/29**	**2/32**	12/20	**2/32**	**1/18**	-

3.3. Discussion

The use of corrected estimates in PETs has been motivated by the small sample anomaly. However, the experimental study provides strong evidence for that this anomaly in fact is of minor importance for random forests of PETs compared to the cost of using the corrected estimates. This cost obviously comes from the bias that is introduced when using any of the corrected estimates, in contrast to when using the relative frequency. Furthermore, the small sample anomaly may actually not be as harmful for random forests of PETs as for single PETs and for some of the alternative ways of generating ensembles of PETs, since the anticipated problem of having singletons (or very few examples) to estimate class probabilities in leaf nodes can be expected to occur less frequently when using a bootstrap replicate of the training examples to grow each tree and all examples to estimate the probabilities, compared to using the same examples to both grow the trees and estimate the probabilities, which by necessity will lead to pure (and therefore small) estimation sets, and compared to using a separate subset of the examples to estimate the probabilities, which hence will contain fewer examples for estimation.

Table 2. Accuracy and AUC

	Accuracy						AUC					
Data set	**AV**	**RF**	**LP**	**M=1**	**M=2**	**M=K**	**AV**	**RF**	**LP**	**M=1**	**M=2**	**M=K**
audiology (24 cl.)	69.50	76.00	68.50	71.50	70.50	49.50	94.67	94.95	94.86	94.57	94.41	93.13
balance-scale (3 cl.)	86.40	86.24	87.36	87.20	87.36	87.52	94.75	94.36	95.27	94.67	94.85	94.96
breast-cancer (2 cl.)	73.43	72.73	73.08	73.43	73.43	73.43	67.44	67.91	68.28	68.07	68.09	68.09
breast-cancer-wisconsin (2 cl.)	96.28	96.42	96.42	96.42	96.42	96.42	99.17	99.20	99.16	99.17	99.15	99.15
car (4 cl.)	96.47	96.99	97.11	97.11	96.93	95.25	99.77	99.84	99.82	99.82	99.81	99.79
cleveland-heart-disease (5 cl.)	56.77	56.77	57.10	57.76	55.12	54.13	80.41	81.05	81.87	81.42	81.55	81.70
crx (2 cl.)	86.79	86.50	86.65	86.65	86.65	86.65	93.03	93.20	93.21	93.19	93.19	93.19
cylinder-bands (2 cl.)	79.44	78.15	78.15	77.96	76.30	76.30	87.54	87.43	86.85	87.06	86.84	86.84
dermatology (6 cl.)	98.09	98.09	98.63	98.09	98.09	97.54	99.92	99.93	99.93	99.93	99.93	99.92
ecoli (8 cl.)	87.20	87.80	87.80	87.50	86.90	85.12	96.85	96.87	97.01	96.94	96.81	96.52
glass (6 cl.)	78.50	78.97	76.64	77.57	77.10	73.36	92.82	93.22	92.45	93.01	92.83	92.29
hepatitis (2 cl.)	85.81	86.45	84.52	85.81	83.23	83.23	85.30	85.59	85.92	85.95	86.03	86.03
house-votes (2 cl.)	96.09	96.55	96.32	96.55	96.32	96.32	99.27	99.32	99.29	99.31	99.30	99.30
image-segmentation (7 cl.)	92.38	93.33	92.38	93.33	93.33	92.38	99.10	99.14	99.06	99.11	99.10	99.06
ionosphere (2 cl.)	94.02	94.02	94.02	94.30	93.73	93.73	98.23	98.15	97.91	98.00	97.89	97.89
iris (3 cl.)	94.67	96.00	95.33	95.33	95.33	95.33	99.04	98.99	99.09	99.03	99.05	99.09
kr-vs-kp (2 cl.)	98.69	98.87	98.75	98.75	98.69	98.69	99.86	99.90	99.91	99.91	99.91	99.91
lung-cancer (3 cl.)	46.88	46.88	46.88	50.00	50.00	50.00	65.54	64.11	64.42	64.07	64.09	64.09
lymphography (4 cl.)	83.78	85.14	84.46	83.78	82.43	82.43	92.95	93.44	93.33	93.41	93.40	93.20
mushroom (2 cl.)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
new-thyroid (3 cl.)	93.95	93.95	93.95	93.49	94.88	94.88	99.35	99.31	99.33	99.31	99.33	99.32
pima-indians-diabetes (2 cl.)	76.43	76.82	76.69	76.56	76.82	76.82	82.45	82.79	83.07	82.94	83.05	83.05
post-operative-patients (3 cl.)	70.00	67.78	67.78	70.00	71.11	71.11	47.15	37.63	36.56	36.02	35.14	34.90
primary-tumor (21 cl.)	46.31	46.90	46.90	47.79	46.31	38.35	79.93	82.63	83.13	82.19	82.14	81.49
promoters (2 cl.)	80.19	87.74	87.74	87.74	87.74	87.74	93.54	95.69	95.44	95.55	95.41	95.41
sick-euthyroid (2 cl.)	97.88	97.91	97.72	97.60	97.44	97.44	97.61	98.28	98.39	98.30	98.30	98.30
soybean-large (19 cl.)	90.23	90.23	85.99	88.27	86.32	81.11	98.95	99.07	99.06	99.07	99.06	98.97
spambase (2 cl.)	94.87	94.96	94.87	94.83	94.74	94.74	98.34	98.60	98.53	98.56	98.53	98.53
spectf (2 cl.)	90.26	91.12	89.97	89.97	88.83	88.83	96.34	96.51	96.07	96.17	95.97	95.97
splice (3 cl.)	96.36	96.49	96.39	96.43	96.46	96.49	99.41	99.42	99.41	99.42	99.42	99.41
tae (3 cl.)	52.32	57.62	54.30	56.29	54.30	55.63	72.19	74.56	72.85	73.68	73.17	72.82
tic-tac-toe (2 cl.)	97.70	98.33	98.12	98.12	98.12	98.12	99.60	99.65	99.63	99.64	99.63	99.63
wine (3 cl.)	98.31	98.31	98.31	98.31	98.31	98.31	99.91	99.91	99.90	99.91	99.91	99.90
yeast (10 cl.)	62.06	62.13	62.06	61.93	61.99	58.42	83.98	84.73	84.44	84.74	84.70	84.43

4. Concluding Remarks

We have investigated the effect of using corrected probability estimates in random forests, and we have compared the use of the (non-corrected) relative class frequency to using the Laplace estimate and the m-estimate. An experiment with 34 datasets from the UCI repository shows that estimating class probabilities using the relative class frequency significantly outperforms or is clearly ahead of both the Laplace estimate and the m-estimate with respect to accuracy, AUC and Brier score. Hence, these results strongly suggest that a non-corrected probability estimate should be used in random forests of PETs, in contrast to what previously has been commonly employed. The experiment further shows that learning random forests of PETs using relative class frequency significantly outperforms learning random forests of classification trees with respect to both accuracy and AUC. However, random forests of classification trees turned out to give clearly better Brier scores compared to random forests of PETs. One direction for future research is to investigate the reasons for this phenomenon. Another direction for future work is to explore calibration methods (i.e., methods that post-process the probabilities output by a model) in conjunction with the different probability estimates [8]. The calibration methods typically do not affect the AUC (since the relative order of examples according to assigned probabilities is not changed), but can have a major impact on accuracy and Brier score.

Acknowledgements

This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104.

References

[1] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1-2):105–139, 1999.

[2] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[3] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(6):1145–1159, 1997.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Table 3. Brier score

Data set	AV	RF	LP	M=1	M=2	M=K
audiology (24 cl.)	0.4060	0.4173	0.7178	0.4549	0.4831	0.6853
balance-scale (3 cl.)	0.2123	0.2199	0.2427	0.2243	0.2318	0.2396
breast-cancer (2 cl.)	0.4128	0.3816	0.3774	0.3761	0.3748	0.3748
breast-cancer-wisconsin (2 cl.)	0.0575	0.0571	0.0605	0.0599	0.0630	0.0630
car (4 cl.)	0.0653	0.0614	0.1046	0.0762	0.0910	0.1171
cleveland-heart-disease (5 cl.)	0.5335	0.5119	0.5332	0.5129	0.5185	0.5351
crx (2 cl.)	0.1918	0.1935	0.2009	0.1971	0.2011	0.2011
cylinder-bands (2 cl.)	0.2917	0.3032	0.3230	0.3151	0.3239	0.3239
dermatology (6 cl.)	0.0501	0.0554	0.1234	0.0684	0.0812	0.1273
ecoli (8 cl.)	0.2104	0.2116	0.3032	0.2223	0.2354	0.3044
glass (6 cl.)	0.3388	0.3451	0.4544	0.3652	0.3834	0.4392
hepatitis (2 cl.)	0.2136	0.2190	0.2301	0.2255	0.2321	0.2321
house-votes (2 cl.)	0.0591	0.0607	0.0690	0.0649	0.0694	0.0694
image-segmentation (7 cl.)	0.1390	0.1435	0.2567	0.1609	0.1783	0.2567
ionosphere (2 cl.)	0.0953	0.1031	0.1144	0.1105	0.1172	0.1172
iris (3 cl.)	0.0686	0.0674	0.0762	0.0691	0.0723	0.0762
kr-vs-kp (2 cl.)	0.0325	0.0338	0.0436	0.0390	0.0437	0.0437
lung-cancer (3 cl.)	0.6158	0.5962	0.6085	0.5993	0.6039	0.6084
lymphography (4 cl.)	0.2392	0.2413	0.3121	0.2564	0.2695	0.2911
mushroom (2 cl.)	0.0001	0.0001	0.0003	0.0002	0.0003	0.0003
new-thyroid (3 cl.)	0.0711	0.0743	0.0885	0.0800	0.0864	0.0930
pima-indians-diabetes (2 cl.)	0.3240	0.3178	0.3180	0.3179	0.3195	0.3195
post-operative-patients (3 cl.)	0.5380	0.4825	0.4862	0.4685	0.4606	0.4553
primary-tumor (21 cl.)	0.7212	0.6752	0.8332	0.6838	0.6942	0.7937
promoters (2 cl.)	0.2874	0.2976	0.3268	0.3142	0.3269	0.3269
sick-euthyroid (2 cl.)	0.0381	0.0397	0.0426	0.0427	0.0454	0.0454
soybean-large (19 cl.)	0.1919	0.1938	0.5934	0.2412	0.2812	0.5836
spambase (2 cl.)	0.0815	0.0830	0.0920	0.0880	0.0924	0.0924
spectf (2 cl.)	0.1486	0.1602	0.1825	0.1755	0.1868	0.1868
splice (3 cl.)	0.0946	0.1030	0.1358	0.1157	0.1272	0.1375
tae (3 cl.)	0.6194	0.5522	0.5839	0.5666	0.5762	0.5836
tic-tac-toe (2 cl.)	0.0919	0.0975	0.1256	0.1139	0.1276	0.1276
wine (3 cl.)	0.0557	0.0604	0.0850	0.0690	0.0770	0.0847
yeast (10 cl.)	0.5331	0.5179	0.6449	0.5246	0.5345	0.5972

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[8] I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. In *Proceedings of the 15th European Machine Learning Conference*. Springer-Verlag, 2004.

[9] T. F. F. Provost and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning*, pages 445–453, 1998.

[10] T. Fawcett. Using rule sets to maximize roc performance. In *Proceedings of the IEEE International Conference on Data Mining*, pages 131–138. IEEE Computer Society, 2001.

[11] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories, Palo Alto, 2003.

[12] C. Ferri, P. Flach, and J. Hernndez-Orallo. Improving the auc of probabilistic estimators trees. In *Proceedings of the 14th European Conference on Artifical Intelligence*, volume 2837, pages 121–132. Springer, 2003.

[13] C. X. Ling and R. J. Yan. Decision tree with better ranking. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 480–487, 2003.

[14] D. Margineantu and T. Dietterich. Improved class probability estimates from decision tree models. Technical report, Department of Computer Science, Oregon State Univ., 2002.

[15] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52(3), 2003.

[16] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.