Calibrating Random Forests

Henrik Boström
Informatics Research Centre
University of Skövde
541 28 Skövde, Sweden
henrik.bostrom@his.se

Abstract

When using the output of classifiers to calculate the expected utility of different alternatives in decision situations, the correctness of predicted class probabilities may be of crucial importance. However, even very accurate classifiers may output class probabilities of rather poor quality. One way of overcoming this problem is by means of calibration, i.e., mapping the original class probabilities to more accurate ones. Previous studies have however indicated that random forests are difficult to calibrate by standard calibration methods. In this work, a novel calibration method is introduced, which is based on a recent finding that probabilities predicted by forests of classification trees have a lower squared error compared to those predicted by forests of probability estimation trees (PETs). The novel calibration method is compared to the two standard methods, Platt scaling and isotonic regression, on 34 datasets from the UCI repository. The experiment shows that random forests of PETs calibrated by the novel method significantly outperform uncalibrated random forests of both PETs and classification trees, as well as random forests calibrated with the two standard methods, with respect to the squared error of predicted class probabilities.

1. Introduction

The ability to output class probability distributions not only allows a learned classifier to be used for classifying or ranking examples (i.e., by assigning to each example the most probable class, or by ordering the examples according to their probability of belonging to a certain class, respectively), but also to be used for calculating the expected utility of different alternatives in decision situations. A classifier that more often predicts the correct class and also generates better rankings than some other classifier may still produce poorer probability estimates, hence possibly leading to decisions of lower utility compared to decisions based on probabilities of the latter. Consequently, correctness of the probability estimates that are output by a classifier may, depending on the task, be of greater importance than both classification accuracy, i.e., the percentage of correctly classified examples, and ranking performance, e.g., as measured by the area under ROC curve (AUC) [10].

It has been observed that classifiers generated by some learning algorithms, despite being both accurate and achieving a high AUC, produce class probabilities that are of rather poor quality, as measured by the squared error of the predicted probabilities, i.e., *Brier score* [7]. One approach to addressing this problem is by means of *calibration*, i.e., learning a function that maps the original probability estimates, or scores, into more accurate probability estimates. Two common calibration methods that have been successfully applied in conjunction with many different learning algorithms, including support-vector machines, boosted decision trees and naïve Bayes, are *Platt scaling* [13] and *isotonic regression* [15]. However, they have turned out to be less effective for random forests [5]. In fact, it has been observed that these methods often require quite large calibration sets, i.e., examples that are used to train the calibration functions, if any improvements are to be observed, and that the use of smaller calibration sets actually may hurt performance [12]. Since random forests are relatively well-calibrated to start with, at least compared to many other methods, this leads to the question of whether or not there is any room for improvement at all in cases when large calibration sets cannot be obtained.

Probability estimation trees (PETs) [14] generalize classification trees [6], in that they have class probability distributions instead of single class labels at the leafs. Like forests of classification trees, forests of PETs have been shown to consistently outperform single PETs [14]. The advantage of using PETs instead of classification trees in forests is however not as obvious as for single trees, since class probability estimates can be obtained also from forests of classification trees, just by calculating the proportion of votes for each class [9]. However, a recent study shows that this strategy in fact results in lower accuracy and area under ROC curve compared to using forests of PETs [3], thus demonstrating the advantage using PETs instead of classification trees also in forests. On the other hand, the same study quite surprisingly reports that forests of classification trees outperform forests of PETs w.r.t. the squared error of the predicted class probabilities. This means that there is indeed some room for improving the predicted class probabilities of forests of PETs. In this work, we will take a closer look at the differences between forests of PETs and classification trees to understand what makes the latter more accurately predict probabilities, and investigate whether this can help devise a more effective calibration method for forests of PETs.

In the next section, we first briefly describe the earlier calibration methods, then continue by investigating the difference between forests of PETs and classification trees, and based on this, propose a novel calibration method. In Section 3, we present an experiment comparing the three calibration methods with respect to Brier score, AUC and accuracy, when used together with both forests of classification trees and PETs. Finally, in Section 4, we summarize the main conclusions from this study and point out some directions for future research.

2. Calibration Methods

In this section, we first describe the calibration task and then briefly recall the two previous calibration methods, *Platt scaling* and *isotonic regression*. Based on a closer examination of the difference between forests of PETs and classification trees, we then introduce a novel calibration method. Finally, we discuss ways of obtaining calibration sets for all three methods.

2.1. The Calibration Task

Given a set of pairs $(\bar{s}_1, \bar{p}_1), \ldots, (\bar{s}_n, \bar{p}_n)$, where $\bar{s}_i = \langle s_{i,1} \ldots s_{i,k} \rangle$ is a probability (or score) vector output by some k-class classifier for some example, and $\bar{p}_i = \langle p_{i,1} \ldots p_{i,k} \rangle$ is the true probability vector for the example, i.e., $p_{i,j} = 1$ if the example belongs to the jth class, and $p_{i,j} = 0$ otherwise, the calibration task concerns finding a mapping from score vectors to probability vectors that minimize the expected loss according to some function. The perhaps most common loss function when evaluating the quality of predicted class probabilities, and which is the one considered also in this study is squared loss, or Brier score [7], i.e., $|\bar{s} - \bar{p}|^2$, where $\bar{s}$ is the vector of predicted probabilities, and $\bar{p}$ is the true probability vector for some example. Other possible loss functions include the negative logarithm of the estimated probability for the correct class, often referred to as *log loss*. In contrast to the former measure, the latter only considers the assigned probability for one (the correct) class. In cases where the classifiers may assign zero probability to one of the classes, and hence a single incorrect such assignment would outweigh all other assignments, more sensitive measures include the squared error of the estimated probability for the correct class [2], as well as the absolute value of this error.

However, the two calibration methods that are described next assume having a binary classification task, with a positive and a negative class, for which the training set becomes $(s_1, p_1), \ldots, (s_n, p_n)$, where s_i is the probability estimate, or score, for that a particular example belongs to the positive class (output by some classifier), and p_i is the true probability of the example belonging to the positive class, which is 1 if the example belongs to the positive class, and 0 otherwise. After describing the two methods, we will discuss how to handle multiclass problems.

2.2. Previous Calibration Methods

2.2.1 Platt scaling

Platt scaling [13] was originally introduced as a method for calibrating support-vector machines. It works by finding the parameters of a sigmoid function maximizing the likelihood of the training set. The function is the following:

$$\hat{p}(c|s) = \frac{1}{1 + e^{As+B}}$$

where $\hat{p}(c|s)$ gives the probability that an example belongs to class c, given that it has obtained the score s, and where A and B are parameters of the function. These are found by gradient descent search, minimizing a particular loss function that was devised in [13]. A computationally more robust version of the gradient descent search algorithm was proposed in [11], which is the one employed in our experiments, together with target probabilities of $\{0, 1\}$ for the training examples rather than the corrected target probabilities suggested in [13], since the former were found to significantly improve the calibration of random forests.

2.2.2 Isotonic regression

In [15], isotonic regression was suggested as a calibration method that can be regarded as general form of binning that does not require any specific number of bins to be predetermined or any limits of the size of each bin. The calibration function, which is assumed to be *isotonic*, i.e., non-decreasing, is a step-wise regression function, which can be learned by an algorithm known as the pair-adjacent violators (PAV) algorithm. Starting with a set of input probability intervals, which borders are the scores in the training set, it works by repeatedly merging adjacent intervals for which the lower interval contains an equally high or higher fraction of examples belonging to the positive class. When eventually no such pair of intervals can be found, the algorithm outputs a function that for each input probability interval returns the fraction of positive examples in the training set in that interval. For a detailed description of the algorithm, see [12].

2.2.3 Handling Multiple Classes

The above two calibration methods suffer from one severe limitation: they are only directly applicable to binary classification tasks. In [16], it is suggested that this problem is addressed by transforming multi-class problems into sets of binary classification tasks, for which binary classifiers and their corresponding calibration functions are generated as well as functions for combining the output of multiple (binary) calibration functions into multi-class probability distributions. In [16], several different options are considered for both transforming the classification task into binary classification tasks as well as for combining the calibration functions. For one-against-all binary classification tasks, i.e., one class is considered to belong to the positive class and all other to the negative, it was observed that more complex strategies for combining multiple calibration functions were not more effective than straightforward normalization, i.e.,

$$\hat{p}(c_i|s_1,\ldots,s_k) = \frac{\hat{p}(c_i|s_i)}{\sum_{j=1}^{k}\hat{p}(c_j|s_j)}$$

where $c_1,\ldots,c_k$ are classes, and $s_1,\ldots,s_k$ are the (uncalibrated) scores from the k binary classifiers. This strategy was therefor recommended in [16], and is also the strategy chosen in our study. It should be noted that for classifiers that, like random forests, are able to output multi-class probability distributions, it is in fact not necessary to learn multiple binary classifiers. Instead, it suffices to learn only the (binary) calibration functions (one for each class) according to the one-against-all strategy, where the scores are obtained from the multi-class classifier. This allows calibration to be performed without changing the way in which the original classifier is generated, and this is the approach also adopted here.

2.3. A Novel Calibration Method

2.3.1 Forests of Classification Trees vs. PETs

In [3], it was quite surprisingly discovered that random forests of classification trees outperform random forests of PETs with respect to squared error of the predicted probabilities, while being outperformed with respect to accuracy and AUC. The only difference between the two methods is the way in which the predicted probabilities are formed. For a forest of PETs, the resulting probability distribution is obtained by averaging the probability distributions output by each PET, while for forests of classification trees, each individual probability distribution is replaced by a more extreme version, assigning a probability of one to the most probable class and zero to all others, before forming the averaged distribution. Hence, this transformation is what results in the improved probability predictions. Figure 1 illustrates the effect of performing this transformation when applied on a random half of the housevotes dataset (where the other half has been used to generate the random forest). It can be seen that probabilities output by the random forest of PETs that are relatively near either 0 or 1 are moved even closer to 0 and 1 respectively, when replacing each probability distribution with the more extreme version, i.e., by which the probability of the most probable class has been maximally adjusted upwards.

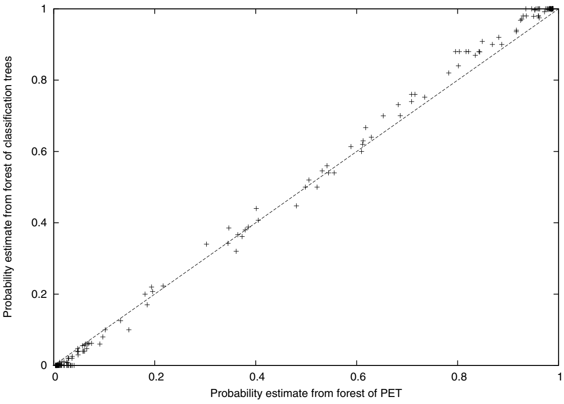

Figure 1. Random forest of classification trees vs. PETs.

2.3.2 Paremeterizing Probability Adjustments

An alternative to correct each individual probability distribution is to instead correct the resulting, averaged, probability distribution. Similarly to the (implicit) correction in random forests of classification trees, we here also consider a correction that increases the estimated probability only for the most probable class, but which rather than pushing the probability to one, parameterizes the increase in the following way:

$$\hat{p}_i = \begin{cases} p_i + r(1-p_i) & \text{if } p_i = Max(\{p_1,\ldots,p_k\}), \\ p_i(1-r) & \text{otherwise} \end{cases}$$

where $p_1,\ldots,p_k$ are the uncorrected probability estimates for the k classes and r, $0 \leq r \leq 1$, is a correction parameter that specifies the percentage of the distance from either one or zero to be reduced. Normalization is not required, since $0 \leq \hat{p}_i \leq 1$ for all $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \hat{p}_i = 1$. Furthermore, since the most probable class for each example will be the same both before and after the correction, the accuracy of the classifier is not affected. Moreover, for binary classification tasks, the AUC is neither affected by this correction, when $r < 1$. To see this, assume the contrary, i.e., there are two examples with probabilities $p_{c,i}$ and $p_{c,j}$ for some class c, such that $p_{c,i} < p_{c,j}$ and $\hat{p}_{c,i} \geq \hat{p}_{c,j}$. This means that $p_{c,i}$ must have been corrected according to case 1 above, while $p_{c,j}$ has been corrected to case 2. Hence, $p_{c,i}$ is the most probable class, which in the two-class case means that $p_{c,i} \geq 0.5$. But since $p_{c,i} < p_{c,j}$, it follows that $p_{c,j} > 0.5$, and hence $p_{c,j}$ could not have been corrected according to case 2, which contradicts the assumption. However, for multi-class tasks, the AUC may indeed be affected. This may happen whenever there is a pair of examples i and j with different true class labels, one being c, and for which the probabilities $p_{c,i}$ and $p_{c,j}$ have the following property: $p_{c,i} < p_{c,j}$ and c is the most probable class for example i, but not for example j.

It should be noted that the above properties not only hold for constant values of r, but also when r is replaced with some non-decreasing (or isotonic) function of the probability that is adjusted upwards. In this study, we consider using a sigmoid function (on the same form as used in Platt scaling), where the parameters A and B of this function are chosen so that the Brier score is minimized for a given calibration set (i.e., pairs of score and true probability vectors).[1]

2.4. Obtaining Calibration Sets for Random Forests

One central question when applying calibration concerns how to obtain the calibration set. One obvious choice is to use the same set of examples both for generating the classifier, and for obtaining scores and true probabilities (by applying the classifier to this set). However, as pointed out in [12], this would lead to that unwanted bias is introduced, e.g., if the classifier is able to perfectly rank the training set, which often is the case for random forests, then the corresponding calibration functions generated by Platt scaling and isotonic regression will just be 0-1 step functions. Another option is to set aside some of the original training examples for the purpose of calibration. However, in cases when there are only limited numbers of training examples, this may have severe effects on both the quality of the learned classifier as well as on the generated calibration function. A computationally more costly approach is to employ n-fold cross-validation, on each iteration using all but one fold to train the classifier, which then is applied on the remaining fold, resulting in that for each example in the training set, there will be a pair in the calibration set with a score that has been obtained from an independently trained classifier. For random forests, and other methods that rely on generating bootstrap replicates, there is another option that involves no additional classifiers being generated, and which therefor is the approach adopted in this study: a score may be obtained for each training example by only averaging the probability distributions from the trees in the forest for which the example is *out-of-the-bag*, i.e., for which the example has not been used in their generation. Since on average about 0.632 of the training examples are used to grow a single tree in the forest, it means that each example will be out-of-bag for a little more than a third of the trees in the forest. Hence, the scores in the calibration set will be based on a sample of votes of the forest, but for large forests, one may expect the scores obtained by sampling to be close to the actual.

3. Empirical Evaluation

In this section, we empirically evaluate the three different calibration methods, Platt scaling, isotonic regression and the novel calibration method, on both forests of classification trees and PETs, with respect to Brier score, accuracy and AUC. We first describe the experimental setup and then present the experimental results.

3.1. Experimental Setup

We consider both random forests of classification trees, as originally proposed in [5], and random forests of probability estimation trees. In this study, each forest consists of 100 unpruned[2] trees, where 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. The size of the subset is in this study equal to the square root of the number of available attributes, as suggested in [5]. The set of examples that is used for estimating class probabilities, i.e., the *estimation examples*, consists of the entire set of training examples.

All compared forests are identical except for the class probability distributions that are used when classifying test instances.[3] For a random forest of classification trees, the probability distribution is formed by averaging the unweighted class votes by the members of the forest, where each member vote for a single (the most probable) class.

[1] Since this loss function differs from the one considered in Platt scaling, the same algorithm for finding the optimal values of A and B cannot be used. We here instead straightforwardly consider values for A and B from the set $\{0, \ldots, 50\}$.

[2] One finding that is common for both forests of classification trees and forests of PETs is that pruning has a detrimental effect [2].

[3] The same examples, random seed, etc. are used for all forests.

Table 1. Ranks for calibrated and uncalibrated random forests.

	CT	CT_I	CT_P	CT_R	PET	PET_I	PET_P	PET_R
Brier score	4.765	4.603	4.647	3.324	5.412	5.632	4.618	3.000
AUC	4.074	6.662	4.515	4.294	3.015	6.779	3.441	3.221
Accuracy		4.118	3.868	3.647		3.809	2.912	2.647

For a random forest of PETs, the probability distribution is obtained by averaging class probability distributions as estimated by the relative class frequencies in the estimation examples.[4] More formally, the probability $p_{k,e,\{t_1,\ldots,t_N\}}$ of an example e belonging to a class k given a forest $\{t_1,\ldots,t_N\}$ is:

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

where $P(t_i,e,k)$ for a classification tree t_i returns 1 if k is the most probable class[5] for example e, and 0 otherwise, and for a PET t_i returns the estimated probability of e belonging to class k according to t_i. For PETs using relative frequencies:

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

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

Random forests of classification trees and PETs are evaluated both without calibration (for which the corresponding classifiers are denoted by CT and PET, respectively) and with the three different calibration methods presented in Section 2: Platt scaling (for which the resulting classifiers are denoted by CT_P and PET_P, respectively), isotonic regression (for which the classifiers are denoted by CT_I and PET_I) and the novel calibration method (for which the classifiers are denoted by CT_R and PET_R).[6]

The eight methods are compared w.r.t. Brier score, AUC and accuracy, using stratified ten-fold cross-validation on the 34 data sets that were considered in [3], which all are taken from the UCI Repository [1]. The average scores obtained for the ten folds are compared.[7]

3.2. Experimental Results

In Table 1, the mean ranks of the eight methods on the 34 datasets are reported with respect to Brier score, AUC and accuracy. Note that since the novel calibration method does not affect accuracy (as shown in Section 2.3), the identical accuracy results obtained for CT and CT_R, as well as PET and PET_R respectively, are not duplicated, which otherwise would bias the accuracy ranks.

When analyzing the results of the methods with respect to the Brier score, a Friedman test [8] shows that the probability of obtaining the average ranks under the null hypothesis, i.e., the choice of method has no impact on Brier score, is 2.36×10^{-5}, hence strongly suggesting that the null hypothesis can be rejected. A post-hoc Bonferroni-Dunn test (assuming that all classifiers are compared to random forests of PETs calibrated with the novel method) shows that all methods, except random forests of classification trees calibrated with the novel method, are significantly outperformed.[8] Hence, the novel calibration method indeed clearly improves the predicted class probabilities of random forest of PETs. This contrasts with isotonic regression that in fact slightly deteriorates the Brier score for random forests of PETs, while Platt scaling results in a limited improvement. Quite surprisingly, the novel calibration method was also able to improve random forests of classification trees.

Regarding the results on AUC for the eight methods on the 34 datasets, the probability of obtaining the average ranks shown in Table 1 under the null hypothesis, i.e., the choice of method has no impact on AUC, is 1.1×10^{-15} according to a Friedman test, hence strongly suggesting that the null hypothesis can be rejected. A post-hoc Nemenyi test (assuming that all 8 classifiers are compared to each other) shows that both methods using isotonic regression are significantly outperformed (at the 0.05 level) by all other methods. None of the other differences are significant according to the Nemenyi test. It can also be seen that the novel calibration method in fact hardly has any effect on the AUC, although it potentially could have had an effect on multi-class problems, as shown in Section 2.3.

Finally, regarding the results on accuracy, the probability of obtaining the average ranks reported in Table 1, if the null hypothesis holds, i.e., the choice of method has no impact

[4] In [3], the use of both the Laplace estimate and the m-estimate were shown to have detrimental effects on accuracy, AUC and Brier score, and are therefor not considered here.

[5] In all cases when there is more than one most probable class, the first in a fixed order of the classes is arbitrarily selected, where the same order is used for all methods.

[6] All methods used the out-of-bag examples as calibration sets, as described in the previous section.

[7] The AUC was calculated according to [10], and for data sets with more than two classes, the total (weighted) AUC is reported.

[8] The differences in ranks are greater than 1.598, which is the critical difference on the 0.05 level, when having eight classifiers and 34 datasets.

on accuracy, is 0.0052 according to a Friedman test, again suggesting that the null hypothesis may be safely rejected. Although random forests of PETs (either uncalibrated or calibrated with the novel method) comes out as a clear winner in terms of accuracy, a post-hoc Nemenyi test (assuming that all 6 classifiers are compared to each other) finds only one of the pairwise differences to be significant at the 0.05 level, namely random forests of PETs vs. random forests of classification trees calibrated with isotonic regression.

4. Concluding Remarks

In this work, we have introduced a novel method for calibrating random forests. The method can be viewed as a parameterized version of the calibration taking place in classification trees, i.e., the probability of the most probable class is moved towards its maximum. An experiment was presented demonstrating that random forests of probability estimation trees calibrated with the novel calibration method significantly outperform uncalibrated random forests with respect to squared error of predicted class probabilities (Brier score). The novel method was furthermore shown to significantly outperform random forests calibrated with either Platt scaling or isotonic regression. The study has also shown that this calibration can be achieved without sacrificing accuracy or area under ROC curve and without requiring large separate calibration sets or changing the way in which the underlying classifier is generated.

Although the proposed method has been developed for random forests in particular, the method is not limited to this type of classifier. One direction for future research is to investigate the effectiveness of the method also for other types of classifier. Another direction for future research concerns investigating alternative functions, as well as algorithms for estimating parameters of these, for determining the degree to which the probabilities should be adjusted.

Acknowledgment

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] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

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

[3] H. Boström. Estimating class probabilities in random forests. In *Proc. of the International Conference on Machine Learning and Applications*, pages 211–216, 2007.

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

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

[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] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[9] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Knowledge Discovery and Data Mining*, pages 155–164, 1999.

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

[11] H-T. Lin, C-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.

[12] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proc. of the 22nd international conference on Machine learning*, pages 625–632, New York, NY, USA, 2005. ACM.

[13] J. Platt. Probabilities for support vector machines. In B. Schoelkopf D. Schuurmans A.J. Smola, P. Bartlett, editor, *Advances in Large Margin Classiers*, pages 61–74. MIT Press, 2000.

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

[15] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proc. 18th International Conference on Machine Learning*, pages 609–616, 2001.

[16] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proc. of the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.