Machine Learning (2019) 108:535–550
https://doi.org/10.1007/s10994-018-5753-x

Efficient Venn predictors using random forests

Ulf Johansson[1] · **Tuve Löfström**[1] · **Henrik Linusson**[2] · **Henrik Boström**[3]

Received: 16 February 2018 / Accepted: 1 August 2018 / Published online: 20 August 2018

Abstract
Successful use of probabilistic classification requires well-calibrated probability estimates, i.e., the predicted class probabilities must correspond to the true probabilities. In addition, a probabilistic classifier must, of course, also be as accurate as possible. In this paper, Venn predictors, and its special case Venn-Abers predictors, are evaluated for probabilistic classification, using random forests as the underlying models. Venn predictors output multiple probabilities for each label, i.e., the predicted label is associated with a probability interval. Since all Venn predictors are valid in the long run, the size of the probability intervals is very important, with tighter intervals being more informative. The standard solution when calibrating a classifier is to employ an additional step, transforming the outputs from a classifier into probability estimates, using a labeled data set not employed for training of the models. For random forests, and other bagged ensembles, it is, however, possible to use the out-of-bag instances for calibration, making all training data available for both model learning and calibration. This procedure has previously been successfully applied to conformal prediction, but was here evaluated for the first time for Venn predictors. The empirical investigation, using 22 publicly available data sets, showed that all four versions of the Venn predictors were better calibrated than both the raw estimates from the random forest, and the standard techniques Platt scaling and isotonic regression. Regarding both informativeness and accuracy, the standard Venn predictor calibrated on out-of-bag instances was the best setup evaluated. Most importantly, calibrating on out-of-bag instances, instead of using a separate calibration set, resulted in tighter intervals and more accurate models on every data set, for both the Venn predictors and the Venn-Abers predictors.

Keywords Probabilistic prediction · Venn predictors · Venn-Abers predictors · Random forests · Out-of-bag calibration

1 Introduction

Many classifiers are able to output not only the predicted class label, but also a probability distribution over the possible classes. Such probabilistic predictions have many obvious uses,

Editor: Lars Carlsson.

✉ Ulf Johansson
ulf.johansson@ju.se

Extended author information available on the last page of the article

one example is to filter out unlikely or very uncertain predictions. Another generic scenario is when the probability estimates are used as the basis for a decision, typically comparing the utility of different options. Naturally, probabilistic prediction requires that the probability estimates are *well-calibrated*, i.e., the predicted class probabilities must reflect the true, underlying probabilities. If this is not the case, the predicted probabilities actually become misleading.

There exist a number of general methods for calibrating probabilistic predictions, but the two most frequently used are *Platt scaling* (Platt 1999) and *isotonic regression* (Zadrozny and Elkan 2001). Both techniques have been successfully applied in conjunction with many different learning algorithms, including support-vector machines, boosted decision trees and naïve Bayes (Niculescu-Mizil and Caruana 2005). However, for single decision trees, as well as bagged trees and random forests, these calibration techniques have turned out to be less effective (Niculescu-Mizil and Caruana 2005), something which partly can be explained by their requirements for large calibration sets. Boström (2008) showed that this problem can be mitigated when employing bagging, e.g., as in the random forest algorithm, by utilizing out-of-bag predictions; in effect allowing all training instances to be used for calibration. In this work, we investigate the use of *Venn predictors* (Vovk et al. 2004), and *Venn-Abers predictors* (Vovk and Petej 2012) as alternative approaches to calibrating probabilities from random forests. Venn predictors (and the special case Venn-Abers predictors) are, under the standard i.i.d. assumption, automatically valid multiprobability predictors, i.e., their probability estimates will be perfectly calibrated, in the long run. The price paid for this rather amazing property is, however, that all probabilistic predictions from a Venn predictor come in the form of intervals.

A formal description of Venn predictors and Venn-Abers predictors is given in Sects. 3 and 3.1, but the overall procedure can be described like this; before the actual prediction, instances are divided into categories. When predicting, we first find the category to which the test instance belongs, and then tentatively classify it as every possible label, one at a time. From this, the frequencies of labels in the chosen category (including the tentative label of the test instance) are used as the estimates of the test label probabilities. Since every possible label is tried, the Venn predictor outputs several (two if it is a two-class problem) probability distributions for the test instance.

Venn predictors can be applied on top of all classifiers, as long as they return not only the class label, but also some score associated with the confidence in that prediction. In this paper, we focus on using the state-of-the-art predictive modeling technique *random forests* (Breiman 2001) as underlying models for Venn predictors. In the empirical investigation, the quality of the probability estimates from the Venn predictors will be compared to both using raw estimates from the random forests and standard calibration techniques. In addition, different versions of Venn predictors will be compared against each other, with regard to accuracy and informativeness. Specifically, since random forests are used as the underlying model, the option to calibrate the estimates on the so called *out-of-bag set*, instead of setting aside a separate data set for calibration, will be investigated.

Previous evaluations of Venn predictors, such as Lambrou et al. (2015), use very few data sets, thus precluding statistical analysis, i.e., they serve mainly as proof-of-concepts. In this paper, we present the first large-scale empirical investigation in which Venn predictors are compared to state-of-the-art methods for calibration of probabilistic predictions, on 22 publicly available data sets.

In the next section, we first define probabilistic prediction and describe random forests, before presenting some standard calibration techniques. The Venn predictors, including the special case of Venn-Abers predictors, are described in Sect. 3. In Sect. 4, we outline the

experimental setup, which is followed by the experimental results presented in Sect. 5. Finally, we summarize the main conclusions in Sect. 6.

2 Background

2.1 Probabilistic prediction

In probabilistic prediction, the task is to predict the probability distribution of the label, given the training set and the test object. The goal is to obtain a *valid* predictor. In general, validity means that the probability distributions from the predictor must perform well against statistical tests based on subsequent observation of the labels. In particular, we are interested in *calibration*, i.e., we want:

$$p(c_j \mid p^{c_j}) = p^{c_j}, \tag{1}$$

where p^{c_j} is the probability estimate for class c_j. It must be noted that validity cannot be achieved for probabilistic prediction in a general sense, see e.g., Gammerman et al. (1998).

2.2 Random forests

A random forest (Breiman 2001) is an ensemble consisting of *random trees*, which are decision trees generated in a specific way to obtain diversity among the trees. Each random tree is trained on a *bootstrap replicate*, i.e., a sample obtained from n training instances by randomly selecting n instances with replacement. This procedure is referred to as *bagging*. Moreover, only a randomly selected subset of the available attributes are considered when choosing each interior split. The instances that were missing in the bootstrap replicate, for a specific tree, are said to be *out-of-bag* for that tree. The random forest algorithm has frequently been demonstrated to achieve state-of-the-art predictive performance, see e.g., Caruana and Niculescu-Mizil (2006) and Delgado et al. (2014) for large-scale comparisons. Random forests can be used for several different tasks, including classification, regression, ranking and probability estimation, see e.g., Boström (2012). In addition to the strong predictive performance, the learning algorithm, being embarrassingly parallel, lends itself to efficient implementation on multi-core platforms, see e.g., Boström (2011) and Jansson et al. (2014).

2.3 Platt scaling

Platt scaling (1999) was originally introduced as a method for calibrating support-vector machines. The method maximizes the likelihood of the training set by finding parameters for the sigmoid function:

$$\hat{p}(c \mid s) = \frac{1}{1 + e^{As+B}}, \tag{2}$$

where $\hat{p}(c \mid s)$ 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, which are found by gradient descent search, minimizing a particular loss function that was devised by Platt (1999).

2.4 Isotonic regression

Zadrozny and Elkan (2001) suggested using isotonic regression for calibrating probabilities. It can be seen as a binning approach, which does not require the number of bins or the bin sizes to be specified. 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. The algorithm repeatedly merges score intervals for which a lower interval has a higher or equal relative frequency of examples labeled as positive, using the original scores as starting points for the process. When eventually no such pair of intervals can be found, the algorithm outputs a function that for each score interval returns the relative frequency of examples with a score in the interval that are labeled as positive. For a detailed description of the algorithm, see Niculescu-Mizil and Caruana (2005).

3 Venn predictors

Venn predictors, as introduced by Vovk et al. (2004), are multi-probabilistic predictors with proven validity properties. The impossibility result mentioned earlier for probabilistic prediction is circumvented in two ways: (i) multiple probabilities for each label are output, with one of them being the valid one; (ii) the statistical tests for validity are restricted to calibration. More specifically, the probabilities must be matched by observed frequencies. As an example, if we make a number of predictions with the probability estimate 0.90, these predictions should be correct in about 90% of the cases.

Venn predictors are related to the more well-known Conformal Prediction (CP) framework, which was introduced as an approach for associating predictions with confidence measures, see e.g. Gammerman et al. (1998) and Saunders et al. (1999). Conformal predictors are applied to the predictions from models built using classical machine learning algorithms, often referred to as the underlying models, and complement the predictions with measures of confidence.

The CP framework produces valid *region predictions*, i.e., the prediction region contains the true target with a pre-defined probability, when examples are drawn according to a fixed underlying distribution. In classification, a region prediction is a (possibly empty) subset of all possible labels.

Similar to CP, Venn predictors were introduced in a transductive setting, which, however, is computationally inefficient, requiring one underlying model to be trained for every label for each new object. Again, as for CP, a more efficient *inductive* version, that requires the training of only one underlying model, has been developed (Lambrou et al. 2015). We now describe inductive Venn predictors, and the concept of multiprobability prediction, following the ideas by Lambrou et al. (2015).

To construct an inductive Venn predictor, the available labeled training examples are split into two parts, the *proper training set*, used to train an underlying model, and a *calibration set* used to estimate label probabilities for each new test example.

Assume we have a training set of the form $\{z_1, \ldots, z_l\}$ where each instance $z_i = (x_i, y_i)$ consists of two parts; an *object* x_i and a *label* y_i. In the inductive setting, this training set is divided into the proper training set $\{z_1, \ldots, z_q\}$ and the calibration set $\{z_{q+1}, \ldots, z_l\}$. When presented with a new test object x_{l+1}, the aim of Venn prediction is to estimate the probability that $y_{l+1} = Y_j$, for each Y_j in the set of possible labels $Y_j \in \{Y_1, \ldots, Y_c\}$. The key idea of

inductive Venn prediction is to divide all calibration examples into a number of *categories* and use the relative frequency of label $Y_j \in \{Y_1, \ldots, Y_c\}$ in each category to estimate label probabilities for test instances falling into that category. The categories are defined using a *Venn taxonomy* and every taxonomy leads to a different Venn predictor. Typically, the taxonomy is based on the underlying model, trained on the proper training set, and for each calibration and test object x_i, the output of this model is used to assign (x_i, y_i) into one of the categories. One basic Venn taxonomy, which can be used with every kind of classification model, simply puts all examples predicted with the same label into the same category.

When estimating label probabilities for a test instance, the category of that instance is first determined using the underlying model, in an identical way as for the calibration instances. Then, the label frequencies of the calibration instances in that category are used to calculate the label probabilities. In addition, again as in CP, the test instance z_{l+1} is included in this calculation. However, since the true label y_{l+1} is not known for the test object x_{l+1}, all possible labels $Y_j \in \{Y_1, \ldots, Y_c\}$ are used to create a set of label probability distributions. Instead of dealing directly with these distributions, an often employed compact representation is to use the lower $L(Y_j)$ and upper $U(Y_j)$ probability estimates for each label Y_j. Let k be the category assigned to the test object x_{l+1} by the Venn taxonomy, and Z_k be the set of calibration instances belonging to category k. Then the lower and upper probability estimates are defined by:

$$L(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}|}{|Z_k| + 1} \tag{3}$$

and:

$$U(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \mid y_m = Y_j\}| + 1}{|Z_k| + 1} \tag{4}$$

In order to make a prediction $\hat{y}_{l+1}$ for x_{l+1} using the lower and upper probability estimates, the following procedure is employed:

$$\hat{y}_{l+1} = \max_{Y_j \in \{Y_1, \ldots, Y_c\}} L(Y_j) \tag{5}$$

The output of a Venn predictor is the above prediction $\hat{y}_{l+1}$ together with the probability interval:

$$[L(\hat{y}_{l+1}), U(\hat{y}_{l+1})] \tag{6}$$

It is proven by Vovk et al. (2005) that the multiprobability predictions produced by Venn predictors are automatically valid, regardless of the taxonomy used. Still, the taxonomy is not unimportant since it will affect both the accuracy of the Venn predictor and the size of the prediction interval. Obviously, smaller probability intervals are more informative, and the probability estimates should preferably be as close to one or zero as possible.

3.1 Venn-Abers predictors

One challenge with Venn predictors is to identify the most suitable taxonomy to use. Venn-Abers predictors (Vovk and Petej 2012) are Venn predictors applicable to two-class problems, where the taxonomy is automatically optimized using isotonic regression. Thus, the Venn-Abers predictor inherits the validity guarantee of Venn predictors.

Many classifiers are *scoring classifiers*, i.e., when they make a prediction for a test object, the output is a *prediction score* $s(x)$. In a two-class problem, with labels 0 and 1, the actual prediction is obtained by comparing the score to a fixed threshold c, and predicting the label of x to be 1 if $s(x) > c$. An alternative to using a fixed threshold c, is to apply an increasing

function g to $s(x)$ to calibrate the scores. After calibration, $g(s(x))$ should be interpreted as the probability that the label for x is 1.

Venn-Abers predictors use isotonic regression, as described in Sect. 2.4, for the calibration. A multiprobabilistic prediction from a Venn-Abers predictor is, in the inductive setting, produced as follows; let s_0 be the scoring function for $\{z_{q+1}, \ldots, z_l, (x_{l+1}, 0)\}$, s_1 be the scoring function for $\{z_{q+1}, \ldots, z_l, (x_{l+1}, 1)\}$, g_0 be the isotonic calibrator for

$$\{(s_0(x_{q+1}), y_{q+1}), \ldots, (s_0(x_l), y_l), (s_0(x_{l+1}), 0)\} \tag{7}$$

and g_1 be the isotonic calibrator for

$$\{(s_1(x_{q+1}), y_{q+1}), \ldots, (s_1(x_l), y_l), (s_1(x_{l+1}), 1)\} \tag{8}$$

Then the probability interval for $y_{l+1} = 1$ is

$$[g_0(s_0(x_{l+1})), g_1(s_1(x_{l+1}))] \tag{9}$$

3.2 Out-of-bag-calibration

Although inductive Venn predictors remedy the computational inefficiencies of their transductive counterparts, by requiring the training of only one underlying model, this typically comes at the expense of informational efficiency, i.e., less accurate models. Since only part of the data can be used to train the underlying model, it will tend to produce less accurate predictions from which the taxonomies are constructed, causing the taxonomies to become less homogeneous with respect to the true class labels. Similarly, the taxonomy categories will contain fewer calibration examples (due to the reduced number of available calibration examples) leading to less fine-grained probability distribution estimates.

Regarding the size of the prediction intervals, previous research has shown that inductive Venn predictors will produce significantly tighter intervals, compared to the transductive approach, see e.g., Lambrou et al. (2015). The reason for this is straightforward; when using the transductive approach, the model is actually re-trained for each new test instance and class, leading to quite unstable models. In the inductive approach, though, the model is both trained and applied to the calibration set only once, i.e., the test instance does not affect the model at all, and only moderately impacts the prediction intervals.

This problem of having to trade informational efficiency for computational efficiency exists also within conformal prediction (Vovk et al. 2005), where a solution has been proposed for scenarios where an ensemble of bagged models is used, see Johansson et al. (2014). Here, calibration is performed using out-of-bag estimation, thus allowing all training data to be used both for training and calibration, without the need to retrain the underlying model for every new test instance. Due to the similarities between Venn prediction and conformal prediction, this out-of-bag calibration technique can easily be extended also to Venn predictors constructed using ensembles of bagged classifiers.

Let OOB_{z_i} be the set of trees in the underlying random forest for which the instance $z_i = (x_i, y_i)$ is out-of-bag, i.e., not included in the bootstrapped training sample. Let $f_{OOB_{z_i}}(x_i)$ be the *out-of-bag prediction* for x_i, i.e., the combined prediction of the ensemble members OOB_{z_i} on the object x_i. We can now assign taxonomies for the training set based on $f_{OOB_{z_i}}(x_i)$, instead of using the underlying random forest. This means that each calibration instance is assigned a category based on a (possibly unique) sub-ensemble, that contains on average about a third of the ensemble members.

In order to retain validity, it is essential that the calibration and test instances are treated equally by the underlying model: in a transductive Venn predictor, all of them are included

in the training set, and in an inductive Venn predictor, none of them are. In out-of-bag calibration, however, all calibration instances are used for training, whereas the test instances are not. Hence, some special considerations need to be made when assigning a category to a test object. One important observation is that, while calibration instances are used during training, the prediction used to assign a category to a calibration instance is always made using a sub-ensemble for which the calibration instance was not used in the bootstrap sample, i.e., the underlying (sub-)models do not have any inherent bias towards the calibration set. Still, the two most straight-forward means of assigning a category to the test instance will not retain exchangeability:

1. Use the entire ensemble (random forest) as the underlying model. This is, perhaps, the most natural choice, since any test instance is by default out-of-bag for all ensemble members. This, however, violates the exchangeability assumption in a fairly obvious way; the full ensemble is expected to provide more accurate predictions than the out-of-bag sub-ensembles used to make predictions for the calibration set, meaning that there is a qualitative difference in the way predictions are made, and hence categories assigned, for the calibration and test instances.
2. Use a randomly selected sub-ensemble (containing approximately one third of the ensemble members). This approach results in a prediction and category assignment for the test object that more closely resembles that of the calibration instances. However, exchangeability is still not guaranteed, as calibration instances are assigned categories based on sub-ensembles where (at most) $l-1$ training examples were used as training data, whereas a randomly selected sub-ensemble will be trained using at most l examples (the full training set). Hence, there is still a small qualitative difference between predictions made for calibration and test instances.

Instead, in order to retain exchangeability between calibration and test instances, we re-use an out-of-bag sub-ensemble from one of the calibration examples. By randomly selecting a calibration example z_r, and using its out-of-bag sub-ensemble OOB_{z_r} to make predictions for the test object x_{l+1}, we ensure that predictions for test objects are qualitatively identical to those made for the calibration instances. This is evident due to the fact that both z_r and z_{l+1} are, by definition, out-of-bag for OOB_{z_r}, and hence, OOB_{z_r} is expected to perform identically on both instances. In total, all predictions for the instances $z_1, \ldots, z_{r-1}, z_{r+1}, \ldots, z_{l+1}$ are made using sub-ensembles of approximately equal size, trained using (at most) $l-1$ examples; in all cases, the prediction made for any z_i is done using a sub-ensemble for which z_i was not included in the training set.

So, during prediction, a random index $r \in [1, l]$ is selected, where l is the size of the training set. A category k is assigned to the test instance x_{l+1} based on $f_{OOB_{z_r}}(x_{l+1})$, i.e., the prediction for x_{l+1} made by the out-of-bag sub-ensemble for z_r. Lower and upper probability estimates are then computed by not including z_r in the calibration set Z_k:

$$L(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \setminus \{z_r\} \mid y_m = Y_j\}|}{|Z_k \setminus \{z_r\}| + 1} \tag{10}$$

and:

$$U(Y_j) = \frac{|\{(x_m, y_m) \in Z_k \setminus \{z_r\} \mid y_m = Y_j\}| + 1}{|Z_k \setminus \{z_r\}| + 1} \tag{11}$$

A full proof of how this procedure maintains exchangeability between calibration and test instances is provided by Boström et al. (2017).

4 Method

In the empirical investigation, we look at different ways of utilizing random forests for probabilistic prediction. All experiments were performed in MatLab, and the random forests were generated using the MatLab implementation of the algorithm, called *TreeBagger*. Here, all parameter values were left at their default values, with the exception of using 300 trees in the forest.

The 22 data sets used are all two-class problems, publicly available from either the UCI repository (Bache and Lichman 2013) or the PROMISE Software Engineering Repository (Shirabad and Menzies 2005). In the experimentation, standard 10×10-fold cross-validation was used.

The taxonomy used for the standard Venn predictors in this study is the label prediction of the underlying model, i.e., all instances predicted with the same label are put into one category. Since all problems are two-class, the resulting taxonomy contains only two categories.

For the actual calibration, we compared using standard Venn predictors and Venn-Abers predictors to Platt scaling and isotonic regression, as well as using no external calibration, i.e., the raw estimates from the forest. In addition, we compared calibrating on the out-of-bag instances to using a separate labeled data set (the *calibration set*) not used for learning the trees. When using a calibration set, $2/3$ of the training instances were used for the tree induction and $1/3$ for the calibration. It must be noted that when calibrating on the out-of-bag instances, we follow the procedure proposed by Boström et al. (2017), as described above, i.e., using a subset of the forest for each prediction to guarantee exchangeability between calibration and test instances. For approaches that employ a separate calibration set, however, the entire forest is used for the predictions. In summary, we compare the following ten approaches:

- **RF-cal**: The raw estimates from the forest estimated from a separate calibration set.
- **RF-oob**: The raw estimates from the forest estimated from the out-of-bag set.
- **Platt-cal**: Standard Platt scaling where the logistic regression model was learned on the calibration set.
- **Platt-oob**: Platt scaling calibrating on the out-of-bag set.
- **Iso-cal**: Standard isotonic regression based on the calibration set.
- **Iso-oob**: Isotonic regression calibrated on the out-of-bag set.
- **VP-cal**: A Venn predictor calibrated on a separate data set and using the predicted label from the underlying model as the category.
- **VP-oob**: A Venn predictor calibrated on the out-of-bag set and using the predicted label from the underlying model as the category.
- **VAP-cal**: A Venn-Abers predictor calibrated on a separate data set.
- **VAP-oob**: A Venn-Abers predictor calibrated on the out-of-bag set.

In the experimentation, we want to evaluate different criteria. For all ten setups, we compare the probability estimates to the true observed accuracies. Specifically, we will evaluate the quality of the probability estimates using the *Brier score* (Brier 1950). For two-class problems, let y_i denote the response variable (class) of instance i, where $y_i = 0$ or 1. Denote the probability estimate that instance i belongs to class 1, by p_i. The Brier Score is then defined as

$$Brier\ Score = \sum_{i=1}^{N} (y_i - p_i)^2, \tag{12}$$

where N is the number of instances. The Bries score is consequently the sum of squares of the difference between the true class and the predicted probability over all instances. The Brier score can be further decomposed into three terms called *uncertainty*, *resolution* and *reliability*. In practice, this is done by dividing the range of probability values, i.e., [0, 1] into a number of K intervals and represent each interval $1, 2, \ldots, K$ by a corresponding typical probability value r_k, see Murphy (1973). Here, the reliability term measures how close the probability estimates are to the true probabilities, i.e., it directly measures how well-calibrated the estimates are. The reliability is defined as

$$Reliability = \frac{1}{N}\sum_{k=1}^{K} n_k(r_k - \phi_k)^2, \tag{13}$$

where n_k is the number of instances in interval k, r_k is the mean probability estimate for the positive class over the instances in interval k, ϕ_k is the proportion of instances actually belonging to the positive class in interval k and N is the total number of instances. It must be noted that the reliability score is defined in the contrary direction compared to the English language, i.e., lower reliability is better. In the experimentation, the number of intervals K was set to 100. When calculating the probability estimate for the positive class from all Venn predictors and Venn-Abers predictors, the center point of the corresponding prediction interval was used. It may be noted that another option for producing a single probability estimate from a Venn predictor prediction interval is suggested by Vovk and Petej (2012). While that method is theoretically sound, providing a regularized value where the estimate is moved towards the neutral value 0.5, the differences between the two methods are most often very small in practice.

For the Venn predictors and the Venn-Abers predictors, we also check the validity by making sure that the observed accuracies, i.e., the percentage of correctly predicted test instances, actually fall in (or at least are close to) the intervals. In addition to the quality of the estimates, there are two additional important metrics when comparing the Venn predictors and the Venn-Abers predictors:

- **Interval size**: The tighter the interval is, the more informative.
- **Accuracy**: The predictive performance of the model is of course vital in all of predictive modeling.

5 Results

We start by investigating the overall quality of the estimates. Table 1 shows the differences between the estimates (averaged over all instances for each data set) and the corresponding accuracies. Looking first at using the raw frequencies from the random forests, we see that while the estimates are fairly accurate, they tend to be too pessimistic. Averaged over all data sets, the difference between the estimates and the actual accuracies is approximately 1.5 percentage points. Using Platt scaling and a separate calibration set produces, on the other hand, too optimistic estimates. Platt scaling on out-of-bag is clearly better, but still systematically optimistic. Interestingly enough, the same holds for isotonic regression, but the estimates are generally worse than Platt scaling. The Venn predictor, though, when looking at these aggregated results, appears to be exceptionally well-calibrated. The Venn-Abers predictor, finally, is rather well-calibrated when using out-of-bag calibration, but too optimistic when calibrated on a separate data set. Even if the differences may appear to be rather small in absolute numbers (approximately 1.5 percentage points on average), the fact is that Platt

Table 1 Quality of estimates

Data set	RF		Platt		Iso		VP		VAP	
	cal	oob	cal	oob	cal	oob	cal	oob	cal	oob
colic	−.062	−.051	.008	.006	.033	.015	−.006	−.002	.030	.010
creditA	−.050	−.046	.005	.002	.018	.007	−.003	−.002	.007	.001
diabetes	.009	.009	.013	.009	.027	.011	−.001	−.001	.009	.000
german	.060	.059	.008	.006	.009	.002	.000	−.001	.027	.011
haber	.107	.117	.023	.020	.046	.014	.007	−.002	.016	−.006
heartC	−.031	−.023	.015	.008	.037	.017	−.007	−.001	.026	.006
heartH	−.005	−.004	.014	.006	.042	.013	−.008	−.004	.017	−.001
heartS	−.036	−.030	.011	.004	.039	.016	−.007	−.001	.026	.002
hepati	−.012	−.016	.024	.010	.060	.028	.003	−.004	.015	−.005
iono	−.046	−.036	.014	.011	.023	.009	−.005	−.002	.012	.000
kc1	.023	.017	−.001	−.004	.015	.008	.006	.000	−.001	−.001
kc2	.039	.036	.018	.009	.038	.013	.009	−.001	.008	−.003
kc3	.015	.015	.019	.017	.027	.017	−.001	.001	−.013	−.009
liver	−.009	−.016	.000	−.014	.042	.017	−.001	−.002	.052	.017
mw	.003	.003	.011	.006	.020	.009	.000	.002	−.018	−.016
pc4	−.017	−.018	.015	.014	.013	.006	−.001	−.001	−.006	−.004
sonar	−.104	−.117	.026	.006	.052	.020	−.012	−.004	.049	.011
spect	−.008	−.011	.004	.001	.013	.005	−.005	−.001	.056	.024
spectf	−.023	−.016	.028	.020	.045	.023	.005	.005	.084	.038
ttt	−.171	−.156	.003	.002	.008	.002	−.003	−.001	.003	−.001
wbc	−.023	−.021	.002	.000	.014	.006	−.007	−.002	.003	−.002
vote	−.028	−.025	.011	.010	.023	.010	−.005	−.002	−.003	−.005
Mean	**−.017**	**−.015**	**.012**	**.007**	**.029**	**.012**	**−.002**	**−.001**	**.018**	**.003**

scaling, isotonic regression and the Venn-Abers predictor turned out to be intrinsically optimistic, while using the random forest outputs was systematically pessimistic. Consequently, one could argue that they all must be considered misleading in this study. The standard Venn predictor, on the other hand, appears to be well-calibrated, specifically there is no inherent tendency to overestimate or underestimate the accuracy. Most importantly, it should be noted that calibrating on out-of-bag improved the quality of the estimates for all setups evaluated.

In order to perform a more detailed analysis, Table 2 shows the reliability scores for the different techniques. As described above, this is a direct measurement of the quality of the probability estimates. The last row shows the average rank for that setup over all data sets.

First it must be noted that in this study, isotonic regression performs clearly worse than using the raw estimates from the random forest. Looking at the mean ranks, Platt scaling is slightly better than using the raw estimates from an out-of-bag calibration set, but clearly worse than calibrating the forest estimates using a separate calibration set. So, based on these results, there is little to gain from using the standard techniques Platt scaling and isotonic regression for calibrating a random forest. Turning to the Venn predictors, however, we see from the mean ranks that all four setups obtained better estimates, i.e., lower reliability scores, compared to the raw estimates from the random forest. Overall, the Venn predictor again showed the most accurate estimates, but when looking at this more detailed level, we

Table 2 Reliability of estimates

Data set	RF		Platt		Iso		VP		VAP	
	cal	oob	cal	oob	cal	oob	cal	oob	cal	oob
colic	.076	.084	.115	.114	.131	.123	.095	.096	.086	.105
creditA	.118	.124	.153	.153	.163	.158	.135	.138	.120	.148
diabetes	.071	.074	.068	.067	.080	.072	.048	.049	.068	.068
german	.029	.028	.006	.005	.014	.011	.002	.001	.010	.010
haber	.057	.058	.012	.008	.030	.019	.007	.003	.030	.017
heartC	.102	.108	.125	.122	.144	.130	.098	.099	.070	.092
heartH	.104	.108	.109	.106	.129	.116	.083	.085	.066	.087
heartS	.098	.104	.123	.120	.143	.130	.098	.101	.070	.093
hepati	.051	.053	.064	.059	.094	.075	.038	.040	.054	.048
iono	.137	.148	.183	.184	.191	.187	.161	.166	.098	.146
kc1	.041	.042	.020	.023	.029	.027	.015	.017	.028	.027
kc2	.077	.078	.056	.053	.076	.063	.036	.035	.058	.062
kc3	.024	.024	.016	.015	.029	.021	.004	.003	.037	.026
liver	.046	.049	.046	.047	.068	.060	.037	.040	.043	.048
mw	.021	.021	.012	.009	.022	.013	.006	.005	.028	.019
pc4	.033	.036	.050	.053	.054	.053	.024	.027	.049	.054
sonar	.057	.065	.135	.142	.155	.152	.092	.109	.050	.073
spect	.015	.014	.006	.003	.013	.007	.002	.001	.004	.003
spectf	.029	.031	.045	.041	.057	.047	.016	.014	.017	.022
ttt	.083	.101	.213	.220	.215	.221	.199	.215	.120	.137
wbc	.196	.199	.211	.211	.219	.215	.197	.203	.132	.170
vote	.088	.093	.105	.108	.117	.113	.083	.087	.060	.079
Mean	**.071**	**.075**	**.085**	**.085**	**.099**	**.092**	**.067**	**.070**	**.059**	**.070**
Rank	**5.10**	**6.33**	**6.05**	**5.81**	**9.33**	**7.90**	**2.57**	**3.05**	**4.05**	**4.81**

see that calibrating on a separate data set was actually slightly better than using the out-of-bag set.

In order to determine whether the observed differences are statistically significant, we used the procedure recommended by Garcıa and Herrera (2008) and performed a Friedman test (Friedman 1937), followed by Bergmann–Hommel's dynamic procedure (Bergmann and Hommel 1988) to establish all pairwise differences. With ten setups and just 22 data sets, only a few differences are actually significant at the $\alpha = .05$ level, see Table 3, where a 'v' shows that the row setup obtained significantly more reliable estimates than the column setup.

Analyzing the different Venn predictors, Table 4 shows the probability intervals, their sizes and the actual accuracies on each data set. An underlined accuracy means that it is outside the prediction interval.

First of all, we see that all setups are valid, i.e., the empirical accuracy is for all setups inside the intervals produced for a very large majority of data sets. Actually, even for the rare data sets where the accuracy is not within the intervals, it is most often very close. While we expect the Venn and the Venn-Abers predictors to be well-calibrated, it must be noted that the intervals produced are much smaller than what is typically the case when using the

Table 3 Statistically significant differences for reliability $\alpha = .05$

Row versus col.	RF-oob	Platt-cal	Platt-oob	Iso-cal	Iso-oob
RF-cal				v	v
RF-oob				v	
Platt-cal				v	
Platt-oob				v	
VP-cal	v	v	v	v	v
VP-oob	v	v		v	v
VAP-cal				v	v
VAP-oob				v	v

original transductive approach, see e.g., Papadopoulos (2013). This is also consistent with the findings by Lambrou et al. (2015).

Comparing the size of the intervals between the different setups, we observe fairly large differences. On average over all data sets, the intervals for VP-oob are smaller than one percentage point (.006), while for VAP-cal it is .075, i.e., over seven percentage points. Looking at the mean ranks, we find that there is a clear ordering, which is the same for every data set; VP-oob produced the smallest intervals, followed by VP-cal, VA-oob and finally VA-cal. Obtaining so tight (and still valid) intervals for the probabilistic predictions is of course a very strong result for the Venn predictor.

Turning to the accuracies, we see that VP-oob is again the best choice. Here, however, VAP-oob is the second best, indicating that the use of an out-of-bag calibration set will result in more accurate models. Interestingly enough, models calibrated on out-of-bag were more accurate than the corresponding models using a separate calibration set for all setups and on every data set. While the fact that using all data for generating the models (which is possible when calibrating on the out-of-bag set) will result in higher accuracies should be no surprise, we must remember that in this setup, the ensemble used for the actual prediction is a subset of the original ensemble. Consequently, the results actually show that a much smaller ensemble (approximately 100 trees) but bagging from all data, will generally be more accurate than a larger ensemble (300 trees) with access to less data (2/3 of the original training set) for the bagging.

For the statistical testing, we again used a Friedman test (Friedman 1937), followed by Bergmann–Hommel's dynamic procedure (Bergmann and Hommel 1988) to establish the pairwise differences, which are shown in Table 5. Here we see that all setups produced significantly higher accuracy than VAP-cal. In addition, VP-oob was significantly more accurate than VP-cal. Looking finally at VP-oob versus VAP-oob, the p value is .07, i.e., while the difference is not significant at $\alpha = .05$, it is still a strong result for VP-oob.

Extending the analysis of the predictive performance, Table 6 also includes the accuracies obtained by the standard setups, i.e., using the raw estimates from the forest, Platt scaling and isotonic regression.

In Table 6, we can make at least two very important observations: (i) every setup using out-of-bag calibration was more accurate than all setups requiring a separate calibration set, and (ii) while the differences are small in absolute numbers, the Venn-predictor calibrated on out-of-bag instances, is actually the most accurate setup overall.

Table 4 Venn predictor and Venn-Abers predictors intervals

Data set	VP								VAP							
	cal				oob				cal				oob			
	Low	High	Size	Acc	Low	High	Size	Acc	Low	High	Size	Acc	Low	High	Size	Acc
colic	.819	.838	.019	.834	.831	.837	.006	.836	.811	.897	.085	.823	.825	.867	.041	.836
creditA	.865	.874	.010	.872	.874	.877	.003	.877	.843	.901	.058	.865	.860	.887	.027	.873
diabetes	.753	.762	.009	.758	.760	.763	.003	.763	.732	.784	.052	.749	.746	.771	.025	.758
german	.699	.706	.007	.702	.702	.704	.002	.704	.712	.745	.034	.701	.707	.722	.015	.703
haber	.709	.732	.023	.713	.715	.723	.008	.721	.653	.725	.072	.673	.675	.708	.033	.698
heartC	.799	.821	.022	.818	.814	.821	.007	.819	.773	.876	.103	.798	.793	.844	.050	.812
heartH	.801	.824	.023	.820	.817	.825	.008	.825	.763	.865	.102	.797	.790	.841	.051	.817
heartS	.800	.824	.024	.819	.818	.826	.008	.823	.770	.879	.109	.798	.795	.849	.053	.820
hepati	.811	.854	.043	.829	.836	.850	.014	.847	.753	.887	.134	.805	.795	.865	.069	.835
iono	.915	.934	.019	.930	.929	.935	.006	.934	.894	.967	.074	.919	.917	.948	.032	.932
kc1	.749	.754	.006	.746	.757	.759	.002	.758	.733	.764	.031	.749	.750	.765	.016	.759
kc2	.772	.790	.018	.772	.780	.786	.006	.784	.738	.816	.078	.769	.767	.803	.037	.788
kc3	.857	.877	.020	.868	.866	.873	.007	.869	.791	.859	.068	.838	.828	.857	.029	.851
liver	.689	.709	.019	.700	.712	.718	.006	.717	.702	.782	.080	.690	.716	.754	.038	.718
mw	.910	.927	.017	.918	.917	.923	.006	.918	.851	.910	.058	.899	.884	.911	.027	.913
pc4	.894	.899	.005	.897	.900	.902	.002	.902	.873	.905	.032	.895	.891	.906	.015	.902
sonar	.780	.811	.032	.807	.825	.835	.011	.834	.775	.905	.129	.791	.812	.881	.069	.835
spect	.872	.894	.022	.888	.883	.892	.009	.888	.902	.988	.086	.890	.897	.930	.033	.890
spectf	.800	.824	.025	.807	.805	.813	.008	.803	.836	.933	.097	.800	.819	.862	.043	.802
ttt	.968	.975	.007	.974	.987	.990	.002	.990	.958	.998	.040	.975	.981	.996	.015	.989
wbc	.937	.951	.014	.951	.949	.954	.005	.953	.913	.971	.058	.939	.935	.961	.026	.950
vote	.854	.867	.013	.865	.866	.870	.004	.870	.814	.884	.070	.852	.842	.879	.036	.865
Mean	**.820**	**.838**	**.018**	**.831**	**.834**	**.840**	**.006**	**.838**	**.800**	**.875**	**.075**	**.819**	**.819**	**.855**	**.035**	**.834**
Rank			**2.00**	**2.67**			**1.00**	**1.43**			**4.00**	**3.81**			**3.00**	**2.10**

Table 5 Statistically significant differences for accuracy $\alpha = .05$

Row versus col.	VAP-cal	VP-cal
VP-oob	v	v
VP-cal	v	
VAP-oob	v	

Table 6 Accuracy for all setups

Data set	RF		Platt		Iso		VP		VAP	
	cal	oob	cal	oob	cal	oob	cal	oob	cal	oob
colic	.834	.836	.834	.835	.825	.834	.834	.836	.823	.836
creditA	.872	.877	.872	.877	.868	.874	.872	.877	.865	.873
diabetes	.759	.763	.759	.762	.752	.759	.758	.763	.749	.758
german	.666	.666	.696	.699	.699	.703	.702	.704	.701	.703
haberman	.674	.666	.705	.703	.690	.708	.713	.721	.673	.698
heartC	.818	.819	.816	.819	.808	.815	.818	.819	.798	.812
heartH	.820	.825	.819	.824	.804	.820	.820	.825	.797	.817
heartS	.819	.823	.818	.822	.810	.822	.819	.823	.798	.820
hepati	.842	.847	.840	.847	.825	.838	.829	.847	.805	.835
iono	.930	.934	.928	.931	.924	.933	.930	.934	.919	.932
kc1	.750	.758	.753	.760	.751	.759	.746	.758	.749	.759
kc2	.781	.784	.786	.792	.778	.791	.772	.784	.769	.788
kc3	.858	.859	.859	.860	.853	.856	.868	.869	.838	.851
liver	.702	.717	.700	.719	.695	.718	.700	.717	.690	.718
mw	.915	.914	.915	.917	.910	.915	.918	.918	.899	.913
pc4	.898	.902	.900	.904	.897	.903	.897	.902	.895	.902
sonar	.807	.834	.811	.842	.803	.838	.807	.834	.791	.835
spect	.885	.888	.887	.889	.885	.889	.888	.888	.890	.890
spectf	.812	.809	.808	.809	.801	.801	.807	.803	.800	.802
ttt	.974	.990	.980	.990	.978	.990	.974	.990	.975	.989
wbc	.951	.953	.950	.953	.947	.952	.951	.953	.939	.950
vote	.865	.870	.863	.868	.858	.867	.865	.870	.852	.865
Mean	**.829**	**.833**	**.832**	**.837**	**.825**	**.836**	**.831**	**.838**	**.819**	**.834**
Mean rank	**6.33**	**3.79**	**6.21**	**2.81**	**8.52**	**4.12**	**6.10**	**2.76**	**9.19**	**5.17**

6 Concluding remarks

This paper has presented the first large-scale comparison of Venn predictors and Venn-Abers predictors to existing techniques for utilizing random forests in probabilistic prediction. Specifically, the novel option to perform the calibration on the out-of-bag instances has been evaluated.

Regarding calibration, as evaluated using the reliability metric, the results show that the standard techniques Platt scaling and isotonic regression were generally ineffective for calibrating a random forest. All four Venn predictors and Venn-Abers predictors, on the other

hand, were better calibrated than both the raw estimates from the random forest, and the standard techniques Platt scaling and isotonic regression.

When comparing the intervals produced by the Venn predictors and the Venn-Abers predictors to the empirical accuracies, it is obvious that all evaluated setups are valid. The interval sizes, however, varied substantially between the different setups. In fact, the ordering between the setups was identical for all data sets; and the best choice was to use a standard Venn predictor calibrated on out-of-bag instances. The intervals produced using that setup were very tight, on average just over .5 percentage points.

Also when considering model accuracy, the best option was to use a standard Venn predictor, and calibrate on out-of-bag instances. That setup was significantly more accurate than both the Venn predictor and the Venn-Abers predictor calibrated on a separate data set. In addition, it was substantially more accurate than Venn-Abers calibrated on out-of-bag.

Generally, it must be noted that calibrating on out-of-bag instead of using a separate calibration set was extremely successful for both Venn predictors and Venn-Abers predictors, resulting in tighter intervals and more accurate models on every data set.

Acknowledgements Funding was provided by "Stiftelsen för Kunskaps- och Kompetensutveckling" (Grant No. 20150185).

References

Bache, K., & Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed 2 Jan 2018.

Bergmann, B., & Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. In *Multiple hypotheses testing*. Springer, pp. 100–115.

Boström, H. (2008). Calibrating random forests. In *IEEE international conference on machine learning and applications*, pp. 121–126.

Boström, H. (2011). Concurrent learning of large-scale random forests. In *Eleventh Scandinavian conference on artificial intelligence, SCAI 2011*, Trondheim, Norway, May 24th–26th, 2011, pp. 20–29.

Boström, H. (2012). Forests of probability estimation trees. *International Journal of Pattern Recognition and Artificial Intelligence*, *26*(2), 2012.

Boström, H., Linusson, H., Löfström, T., & Johansson, U. (2017). Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, *81*(1–2), 125–144.

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

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Machine learning, proceedings of the twenty-third international conference (ICML 2006)*, Pittsburgh, Pennsylvania, USA, June 25–29, 2006, pp. 161–168.

Delgado, M. F., Cernadas, E., Barro, S., & Amorim, D. G. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(1), 3133–3181.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of American Statistical Association*, *32*, 675–701.

Gammerman, A., Vovk, V., & Vapnik, V. (1998). Learning by transduction. In *Proceedings of the fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann, pp. 148–155.

Garcıa, S., & Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, *9*(2677–2694), 66.

Jansson, K., Sundell, H., Boström, & H. (2014). gpurf and gpuert: Efficient and scalable GPU algorithms for decision tree ensembles. In *2014 IEEE international parallel & distributed processing symposium workshops*, Phoenix, AZ, USA, May 19–23, 2014, pp. 1612–1621.

Johansson, U., Boström, H., Löfström, T., & Linusson, H. (2014). Regression conformal prediction with random forests. *Machine Learning*, *97*(1–2), 155–176. ISSN 0885-6125.

Lambrou, A., Nouretdinov, I., & Papadopoulos, H. (2015). Inductive venn prediction. *Annals of Mathematics and Artificial Intelligence*, *74*(1), 181–201.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*(4), 595–600.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning*. ACM, pp. 625–632.

Papadopoulos, H. (2013). Reliable probabilistic classification with neural networks. *Neurocomputing*, *107*(Supplement C), 59–68.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*. MIT Press, pp. 61–74.

Saunders, C., Gammerman, A., & Vovk, V. (1999). Transduction with confidence and credibility. In *Proceedings of the sixteenth international joint conference on artificial intelligence (IJCAI'99)*, Vol. 2, pp. 722–726.

Shirabad, J. S., & Menzies, T. J. (2005). *The PROMISE repository of software engineering databases*. School of Information Technology and Engineering, University of Ottawa, Canada. http://promise.site.uottawa.ca/SERepository. Accessed 2 Jan 2018.

Vovk, V., & Petej, I. (2012). Venn-abers predictors. arXiv preprint arXiv:1211.0025.

Vovk, V., Shafer, G., & Nouretdinov, I. (2004). Self-calibrating probability forecasting. In *Advances in neural information processing systems*, pp. 1133–1140.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th international conference on machine learning*, pp. 609–616.

Affiliations

Ulf Johansson[1] · Tuve Löfström[1] · Henrik Linusson[2] · Henrik Boström[3]

Tuve Löfström
tuve.lofstrom@ju.se

Henrik Linusson
henrik.linusson@hb.se

Henrik Boström
bostromh@kth.se

1 Department of Computer Science and Informatics, Jönköping University, Jönköping, Sweden

2 Department of Information Technology, University of Borås, Borås, Sweden

3 School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden