On the Calibration of Aggregated Conformal Predictors

Henrik Linusson HENRIK.LINUSSON@HB.SE
Dept. of Information Technology, University of Borås, Sweden

Ulf Norinder ULF.NORINDER@SWETOX.SE
Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Sweden
Dept. of Computer and Systems Sciences, Stockholm University, Sweden

Henrik Boström HENRIK.BOSTROM@DSV.SU.SE
Dept. of Computer and Systems Sciences, Stockholm University, Sweden

Ulf Johansson ULF.JOHANSSON@JU.SE
Tuve Löfström TUWE.LOFSTROM@JU.SE
Dept. of Computer Science and Informatics, Jönköping University, Sweden
Dept. of Information Technology, University of Borås, Sweden

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Conformal prediction is a learning framework that produces models that associate with each of their predictions a measure of statistically valid confidence. These models are typically constructed on top of traditional machine learning algorithms. An important result of conformal prediction theory is that the models produced are provably valid under relatively weak assumptions—in particular, their validity is independent of the specific underlying learning algorithm on which they are based. Since validity is automatic, much research on conformal predictors has been focused on improving their informational and computational efficiency. As part of the efforts in constructing efficient conformal predictors, aggregated conformal predictors were developed, drawing inspiration from the field of classification and regression ensembles. Unlike early definitions of conformal prediction procedures, the validity of aggregated conformal predictors is not fully understood—while it has been shown that they might attain empirical exact validity under certain circumstances, their theoretical validity is conditional on additional assumptions that require further clarification. In this paper, we show why validity is not automatic for aggregated conformal predictors, and provide a revised definition of aggregated conformal predictors that gains approximate validity conditional on properties of the underlying learning algorithm.

Keywords: Confidence Predictions, Conformal Prediction, Classification, Ensembles

1. Introduction

Conformal predictors (Gammerman et al., 1998; Gammerman and Vovk, 2007; Vovk et al., 2006) are predictive models, e.g., classifiers or regression models, that output predictions with a measure of statistically valid confidence. Given a test object, a conformal predictor outputs a multi-valued prediction (i.e., a set or an interval) that contains the true output value with a user-specified predefined probability. This property of statistical validity requires only that the training examples and test objects are *exchangeable*—a requirement that is weaker than the common *i.i.d.* assumption.

Conformal predictors are very flexible in that we can construct them on top of any traditional classification or regression algorithm. Formally, we define a so-called *nonconformity measure*, that ranks possible labels of a test object according to their level of (dis-)agreement with an observed distribution, and such nonconformity measures are typically based on traditional machine learning methods. The nonconformity measure chosen does not affect the validity of the conformal predictor, however, our choice of underlying model may affect the *informational efficiency* of the conformal predictor (in essence, the size of the predictions it outputs). There exists a natural confidence-efficiency trade-off, such that predictions necessarily grow larger when the user expects a greater level of confidence, however, different instantiations of conformal predictors (using different nonconformity measures) may differ in terms of informational efficiency even when applied to the same learning problem, at the same confidence level. Hence, much effort has been spent in assessing the informational efficiency of conformal predictors utilizing nonconformity measures based on different kinds of machine learning algorithms, e.g., support vector machines (Gammerman et al., 1998; Saunders et al., 1999; Toccaceli et al., 2016), ridge regression (Burnaev and Vovk, 2014), neural networks (Papadopoulos, 2008; Papadopoulos and Haralambous, 2011; Löfström et al., 2013; Johansson et al., 2015), random forests (Devetyarov and Nouretdinov, 2010; Bhattacharyya, 2013; Johansson et al., 2014; Boström et al., 2017), decision trees (Johansson et al., 2013) and k-nearest neighbors (Papadopoulos et al., 2011).

Due to the fact that conformal predictors exist on top of standard machine learning methods, *computational efficiency* is also of concern. Early specifications of conformal predictors (Gammerman et al., 1998) defines them in a *transductive* manner, where the underlying model must be retrained each time a new test object is obtained. The intractability of transductively computing prediction regions for a sequence of test objects motivated the development of *inductive* conformal predictors, that require only that the underlying model is trained once (Papadopoulos et al., 2002; Vovk et al., 2006; Papadopoulos, 2008; Vovk, 2013). A significant drawback of inductive conformal predictors, however, is that they require some training examples to be left out from training the underlying predictor, and instead set aside for calibration of the conformal predictor. This is in contrast to transductive conformal predictors, where all available training data can be used for both training and calibration, and leads to inductive conformal predictors having a lower informational efficiency than transductive versions on finite sequences, particularly when the total amount of available training data is relatively small.

As such, not only might a user of conformal prediction need to trade-off confidence for informational efficiency, but also informational efficiency for computational efficiency. A suggested solution for this dilemma are a kind of conformal predictor ensembles, proposed by Vovk (2015) as *cross-conformal predictors* and generalized by Carlsson et al. (2014) as *aggregated conformal predictors*. Here, several underlying models are constructed, each time leaving out a different subset of the training data using a suitable resampling method (e.g., cross-validation, bootstrap sampling or random subsampling), so that training examples may be used for both training and calibration (albeit for different members of the conformal predictor ensemble). This procedure has been shown to be able to improve informational efficiency compared to inductive conformal prediction, while maintaining a relatively low computational cost (Vovk, 2015; Carlsson et al., 2014). However, in contrast to transductive and inductive conformal predictors, aggregated conformal predictors have not been shown

to obtain automatic validity—at least not without imposing additional requirements that are not yet fully understood (Vovk, 2015; Carlsson et al., 2014).

This paper aims to provide a comprehensive analysis of aggregated conformal predictors, in order to ascertain under what circumstances—if any—we can consider them valid conformal predictors.

2. Conformal Prediction

Given a test object $\boldsymbol{x}_{n+1} \in X$ and a user-specified *significance level* $\epsilon \in (0, 1)$ a conformal classifier outputs a *prediction set* $\Gamma^{\epsilon}_{n+1} \subseteq Y$ that contains the true output value $y_{n+1} \in Y$ with *confidence* $1 - \epsilon$ (Vovk et al., 2006).

In order to output such prediction sets, conformal predictors utilize a *nonconformity function* $f : Z^* \times Z \to \mathbb{R}$, where $Z : X \times Y$, and $\alpha_i = f(\zeta, z_i)$ is a measure of the nonconformity (we can think of nonconformity as strangeness, unlikelihood or disagreement with respect to a particular problem space) of an object x_i and label y_i (together referred to as a *pattern*) $z_i = (x_i, y_i) \in Z$ in relation to the sequence $\zeta \in Z^*$. Conformal predictors are automatically well-calibrated regardless of the choice of f, but in order to produce informationally efficient (i.e., small) prediction sets, it is necessary that f is able to rank patterns based on their apparent strangeness sufficiently well. As such, a standard method of defining a nonconformity function is to base it on a traditional machine learning model, according to

$$f\left(\zeta, (\boldsymbol{x}_i, y_i)\right) = \Delta\left(h\left(\boldsymbol{x}_i\right), y_i\right), \tag{1}$$

where h is a predictive model—often referred to as the *underlying model* of the conformal predictor—trained on the sequence ζ, and Δ is some function that measures the prediction errors of h. Intuitively, the prediction error for nonconforming (uncommon) patterns $(\boldsymbol{x}_i, y_i)$ will be large (since, if they are uncommon, h will not have seen many similar training examples), and thus, such patterns are assigned larger nonconformity scores than more common patterns.

Given a sequence of training examples, $Z^n = \{z_1, \ldots, z_n\}$, a test object $\boldsymbol{x}_{n+1}$, and a tentative test label $\tilde{y} \in Y$, we construct the extended sequence $Z^{n+1} = Z^n \cup \{(\boldsymbol{x}_{n+1}, \tilde{y})\}$. We then compute the nonconformity scores of the training patterns $z_i \in Z^n$ as

$$\alpha^{\tilde{y}}_i = f\left(Z^{n+1} \setminus z_i, z_i\right), \tag{2}$$

and the nonconformity score for the tentatively labeled test pattern as

$$\alpha^{\tilde{y}}_{n+1} = f\left(Z^{n+1} \setminus z_{n+1} = Z^n, (\boldsymbol{x}_{n+1}, \tilde{y})\right). \tag{3}$$

The corresponding (smoothed) conformal predictor is then defined as the set predictor

$$\Gamma^{\epsilon}_{n+1} = \left\{\tilde{y} \in Y : p^{\tilde{y}}_{n+1} > \epsilon\right\}, \tag{4}$$

where

$$p^{\tilde{y}}_{n+1} = \frac{\left|\left\{z_i \in Z^{n+1} : \alpha^{\tilde{y}}_i > \alpha^{\tilde{y}}_{n+1}\right\}\right| + \theta_{n+1}\left|\left\{z_i \in Z^{n+1} : \alpha^{\tilde{y}}_i = \alpha^{\tilde{y}}_{n+1}\right\}\right|}{n+1}, \tag{5}$$

where $\theta_{n+1} \sim U[0,1]$. If the sequence $\{z_1, \ldots, z_{n+1}\}$ is *exchangeable*, the probability of making an erroneous prediction, i.e., excluding the true target label y_{n+1}, is asymptotically ϵ as $\lim_{n\to\infty}$.

Definition 1 (Exchangeability) *A sequence $\{z_1, \ldots, z_{n+1}\}$ is exchangeable if the joint probability distribution $P(\{z_1, \ldots, z_{n+1}\}) = P\left(\{z_{\pi(1)}, \ldots, z_{\pi(n+1)}\}\right)$ is invariant under any permutation π on the set of indices $i = 1, \ldots, n+1$, i.e., all orderings of the observations $z_1, \ldots, z_{n+1}$ are equiprobable. Exchangeable sequences can be obtained through sampling observations (with or without replacement) from stationary processes, e.g., drawing numbers $x \in \mathbb{Z}$ according to a fixed, arbitrary, probability distribution $X \sim \mathbb{P}$.*

2.1. Inductive Conformal Prediction

In the definition of conformal predictors given in the previous section, we calculate the nonconformity $\alpha_i^{\tilde{y}}$ of a pattern $z_i = (\boldsymbol{x}_i, y_i) \in Z^{n+1}$, where $Z^{n+1} = Z^n \cup \{(\boldsymbol{x}_{n+1}, \tilde{y})\}$, relative to the sequence $Z^{n+1} \setminus z_i$; see Equations (2) and (3). This has two important consequences. First, we note that the nonconformity of any training example $z_i \in Z^n$ is dependent on the specifics of the tentatively labeled test pattern $(\boldsymbol{x}_{n+1}, \tilde{y})$, meaning the nonconformity scores for all training examples must be recomputed when either $\boldsymbol{x}_{n+1}$ or $\tilde{y}$ changes. Consequently, these nonconformity scores must be recomputed $|Y|$ times for each test object (once for every possible value of $\tilde{y}$). Second, we note that each nonconformity score $\alpha_i^{\tilde{y}} \in \alpha_1^{\tilde{y}}, \ldots, \alpha_{n+1}^{\tilde{y}}$ is computed from a unique sequence $\zeta_i \subset Z^{n+1}$. If f is dependent on an underlying machine learning model h, see Equation (1), this means that h must be retrained once for every pattern $z_1, \ldots, (\boldsymbol{x}_{n+1}, \tilde{y})$, for every specific test pattern $(\boldsymbol{x}_{n+1}, \tilde{y})$. In total, this process of *transductive* conformal prediction (TCP) requires that the underlying model h is retrained $(n+1)|Y|$ times for each test object $\boldsymbol{x}_{n+1}$, which incurs a very large computational cost when the computation of h is non-trivial.

It is possible to reduce the computational complexity by simply computing the nonconformity scores $\alpha_1^{\tilde{y}}, \ldots, \alpha_{n+1}^{\tilde{y}}$ from a common sequence Z^{n+1} as

$$\alpha_i^{\tilde{y}} = f\left(Z^{n+1}, z_i\right), \tag{6}$$

where $z_i \in Z^{n+1}$, however, this still requires that the model is recomputed $|Y|$ times for each test object. Additionally, the informational efficiency of a conformal predictors defined using Equation (6) might suffer when the underlying model is unstable, i.e., the learning algorithm is highly variant with respect to the specific example patterns used during training (Linusson et al., 2014).

An alternative approach is to define an *inductive* conformal predictor (Papadopoulos et al., 2002; Vovk et al., 2006), *ICP*, where the underlying model only needs to be computed once. Here, the training set Z^n is divided into two non-empty disjoint subsets: the *proper training set* Z^t and the *calibration set* Z^c. The underlying model h is inferred from the training examples in Z^t, and nonconformity scores are computed for the calibration set and the test pattern (but not the proper training set), as

$$\alpha_i = f\left(Z^t, z_i\right), \tag{7}$$

where $z_i \in Z^c$, and

$$\alpha_{n+1}^{\tilde{y}} = f\left(Z^t, (\boldsymbol{x}_{n+1}, \tilde{y})\right), \tag{8}$$

respectively.

The p-value for a test object $(\boldsymbol{x}_{n+1}, \tilde{y})$ is then defined as

$$p_{n+1}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z^c : \alpha_i > \alpha_{n+1}^{\tilde{y}}\right\}\right| + \theta_{n+1}\left(\left|\left\{z_i \in Z^c : \alpha_i = \alpha_{n+1}^{\tilde{y}}\right\}\right| + 1\right)}{c+1}, \tag{9}$$

i.e., the p-value of a test pattern is calculated only from the nonconformity scores of the calibration examples and the test pattern itself. Since the nonconformity scores of examples in the calibration set are independent of the test pattern (regardless of the label being tested), only $\alpha_{n+1}^{\tilde{y}}$ needs to be updated during prediction.

While inductive conformal predictors are much more efficient than transductive conformal predictors in a computational sense, their informational efficiency is typically lower, since only part of the data can be used for training and calibration respectively (this difference is accentuated in particular when the available data is small).

2.2. Cross-Conformal Predictors

As a means to alleviate the computational inefficiency of transductive conformal predictors, and the (relative) informational inefficiency of inductive conformal predictors, *cross-conformal predictors* (*CCP*) were developed by Vovk (2015). Here, the training set Z^n is divided into k non-empty disjoint subsets, $Z_1, \ldots, Z_k$, and a predictive model h_l is induced from each set $Z_{-l} = \cup_{r=1,\ldots,k} Z_r \setminus Z_l$ (much like the well-known cross-validation method). Nonconformity scores are computed for the calibration examples in each fold using h_l as

$$\alpha_{i,l} = f\left(Z_{-l}, z_i\right), \tag{10}$$

where $z_i \in Z_l$. For the test pattern $(\boldsymbol{x}_{n+1}, \tilde{y})$, k separate nonconformity scores are obtained according to

$$\alpha_{n+1,l}^{\tilde{y}} = f\left(Z_{-l}, (\boldsymbol{x}_{n+1}, \tilde{y})\right), \tag{11}$$

where $l = 1, \ldots, k$, and the corresponding p-value is then calculated as

$$p_{n+1}^{\tilde{y}} = \frac{\sum_{l=1}^{k}\left[\left|\left\{z_i \in Z_l : \alpha_{i,l} > \alpha_{n+1,l}^{\tilde{y}}\right\}\right| + \theta_{n+1,l}\left(\left|\left\{z_i \in Z_l : \alpha_{i,l} = \alpha_{n+1,l}^{\tilde{y}}\right\}\right|\right)\right] + \theta_{n+1}}{n+1}. \tag{12}$$

As noted by Vovk (2015), if a separate p-value is defined for each fold as

$$p_{n+1,l}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z_l : \alpha_{i,l} > \alpha_{n+1,l}^{\tilde{y}}\right\}\right| + \theta_{n+1,l}\left(\left|\left\{z_i \in Z_l : \alpha_{i,l} = \alpha_{n+1,l}^{\tilde{y}}\right\}\right| + 1\right)}{|Z_l| + 1}, \tag{13}$$

then

$$p_{n+1}^{\tilde{y}} = \bar{p}_{n+1}^{\tilde{y}} + \frac{k-1}{n+1}\left(\bar{p}_{n+1}^{\tilde{y}} - 1\right) \approx \bar{p}_{n+1}^{\tilde{y}}, \tag{14}$$

where $\bar{p}_{n+1}^{\tilde{y}} = \frac{1}{k}\sum_{l=1}^{k} p_{n+1,l}^{\tilde{y}}$, given that $k \ll n$.

Papadopoulos (2015) provides further details on constructing cross-conformal predictors for regression problems.

2.3. Bootstrap Conformal Predictors

Also introduced by Vovk (2015) are *bootstrap conformal predictors* (*BCP*), where the underlying models $h_1, \ldots, h_k$ are trained using a bootstrap sampling procedure. Here, a set of samples $Z_{-1}, \ldots, Z_{-k}$ are drawn (with replacement) from Z^n, each sample of size n. For each bootstrap sample, an underlying model is induced, and nonconformity scores are computed using a particular model h_l analogously to cross-conformal predictors, i.e.,

$$\alpha_{i,l} = f\left(Z_{-l}, z_i\right), \tag{15}$$

where $z_i \in Z_l$ and $Z_l = Z^n \setminus Z_{-l}$, and

$$\alpha^{\tilde{y}}_{n+1,l} = f\left(Z_{-l}, (\boldsymbol{x}_{n+1}, \tilde{y})\right), \tag{16}$$

where $l = 1, \ldots, k$. The p-values are then defined by

$$p^{\tilde{y}}_{n+1} = \frac{\sum_{l=1}^{k} \left[\left| \left\{ z_i \in Z_l : \alpha_{i,l} > \alpha^{\tilde{y}}_{n+1,l} \right\} \right| + \theta_{n+1,l} \left(\left| \left\{ z_i \in Z_l : \alpha_{i,l} = \alpha^{\tilde{y}}_{n+1,l} \right\} \right| \right) \right] + \frac{t}{n}\theta_{n+1}}{t + \frac{t}{n}}, \tag{17}$$

where t is the total size of the calibration sets, i.e., $\sum_{l=1}^{k} |Z_l|$.

2.4. Aggregated Conformal Predictors

Carlsson et al. (2014) provide a generalization of conformal predictors constructed from multiple inductive conformal predictors (e.g., cross-conformal predictors and bootstrap conformal predictors), dubbed *aggregated conformal predictors* (*ACP*).

Given a collection of k proper training sets $Z_{-1}, \ldots, Z_{-k}$ and their complementary calibration sets $Z_1, \ldots, Z_k$, nonconformity functions for the calibration examples and test patterns are defined in the same manner as bootstrap conformal predictors (and cross-conformal predictors); see Equations (15) and (16), respectively.

The p-values are defined as

$$p^{\tilde{y}}_{n+1} = \frac{1}{k} \sum p^{\tilde{y}}_{n+1,l}, \tag{18}$$

using the same definition of $p^{\tilde{y}}_{n+1,l}$ as given in Equation (13).

The definition of ACP thus closely resembles that of CCP, in particular when we take into consideration Equation (14), however, here we are not explicitly bound by some particular sampling scheme in constructing the k calibration sets (e.g., cross-validation or bootstrap sampling). Instead, the definition of ACP puts a more general constraint on the procedure of constructing calibration sets Z_l (and their corresponding training sets Z_{-l}), called *consistent resampling* (Carlsson et al., 2014, Definition 1-2). For completeness, we restate these definitions here.

Definition 2 (Exchangeable resampling) *Let $Z^{n+1} = \{z_1, \ldots, z_{n+1}\}$ be a sequence of examples drawn from the problem space $Z \sim \mathbb{P}$, and let $Z^* = \{z^*_1, \ldots, z^*_m\} \subseteq Z^{n+1}$ be a sequence resampled from Z^{n+1}. We call this resampling exchangeable if*

$$P\left(\{z_1, \ldots, z_m\}\right) = P\left(\left\{z_{\pi(1)}, \ldots, z_{\pi(m)}\right\}\right),$$

for any permutation π of the indices $1, \ldots, m$.

Definition 3 (Consistent resampling) *Let $T = T(z_1, \ldots, z_{n+1}, \mathbb{P})$ be a statistic and $T^* = (z_1^*, \ldots, z_m^*, \mathbb{P}_{n+1})$ be an exchangeably resampled version of T. Further, let G_{n+1} and G_{n+1}^* be the probability distributions of T and T^*, respectively. We call this resampling consistent (with respect to T) if*

$$\sup_z \left| G_{n+1} - G_{n+1}^* \right| \to 0 \text{ as } n \to \infty \text{ and } m \to \infty \,.$$

Carlsson et al. (2014, Proposition 1) finally conclude that an ACP is valid when the calibration sets $Z_1, \ldots, Z_k$ are consistently resampled from Z^n with respect to α_t, where α_t is

$$\underset{\alpha_t \in Z^c}{\operatorname{argmax}} \frac{|\{z_i \in Z^c : \alpha_i \geq \alpha_t\}| + 1}{c+1} > \epsilon, \tag{19}$$

i.e., the threshold nonconformity score defining the border p-value $p_{t+1} < \epsilon < p_t < p_{t-1}$. (Arguably, since $p \sim U[0,1]$ is discrete for a finite c in a non-smoothed inductive conformal predictor, averaging a repeated sampling of $p_t^* > \epsilon$ might provide us with a smoother decision border, given that the sampling is performed with care. Consult Carlsson et al. (2014, Section 2.1) for a more detailed discussion.)

We note here that while Carlsson et al. (2014) state that a consistent resampling of the calibration set is a sufficient criterion for obtaining valid aggregate p-values, no prescriptions are provided as to how such a consistent resampling might be obtained. We will return to this line of thought in Section 3.2.

3. Calibration of Conformal Predictors

A key insight regarding conformal predictors regards the distribution of the p-values generated within the process. We can express two particularly interesting criteria regarding these p-values (Vovk et al., 2006):

I If a sequence $z_1, \ldots, z_{n+1}$ is exchangeable, then $p_i^{y_i} \sim U[0,1]$, and

II Criterion **I** is not dependent on the choice of f.

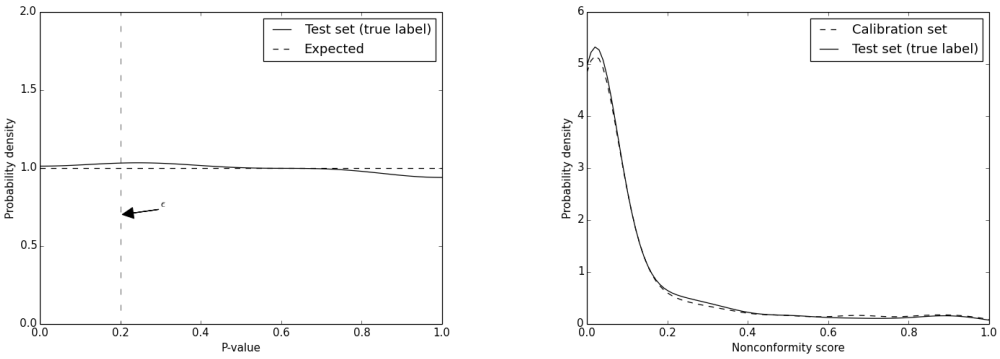
The first criterion is a necessary condition for conformal predictors to be well-calibrated, i.e., make errors at a frequency of exactly ϵ. A conformal predictor rejects any label $\tilde{y}$ for which $p_{n+1}^{\tilde{y}} \leq \epsilon$, hence, in order to make errors at a frequency of ϵ, it must hold that

$$\lim_{n \to \infty} P\left(p_{n+1}^{y_{n+1}} \leq \epsilon\right) = \epsilon \,, \tag{20}$$

for an exactly calibrated conformal predictor. As illustrated in Figure 1(*a*), this becomes true exactly when the p-values are uniformly distributed (whenever we are testing the true output label y_{n+1}) since $\int_0^\epsilon p / \int_0^1 p = \epsilon$. To provide some further intuition regarding criterion **I**, we can restate it in two different manners:

1. Given two exchangeable sequences of examples—a calibration set Z^c, and a test set Z^r—the nonconformity scores $\alpha_r : z_r \in Z^r$ are distributed identically to the nonconformity scores $\alpha_c : z_c \in Z^c$. This is illustrated in Figure 1(*b*).

2. Let Z^{n+1} be an exchangeable sequence of examples, and $\alpha_1, \ldots, \alpha_{n+1}$ be the nonconformity scores computed from $z_1, \ldots, z_{n+1}$. Let $z_1, \ldots, z_n$ be the calibration set examples, and $\pi(1), \ldots, \pi(n)$ denote a permutation of the indices such that $\alpha_{\pi(i)} \leq \alpha_{\pi(i+1)}$. If α_{n+1} is the nonconformity score of the test pattern, z_{n+1}, then all values of $\pi(n+1) \in \{1, \ldots, n+1\}$ are equiprobable, unconditional on z_{n+1}. We can view this in terms of a ranking problem: if we rank each pattern $z_1, \ldots, z_{n+1}$ (using the nonconformity measure as our ranking function), then all ranks $1, \ldots, n+1$ are equally likely for the test pattern z_{n+1}.



(*a*) P-value distribution (*b*) Nonconformity score distribution

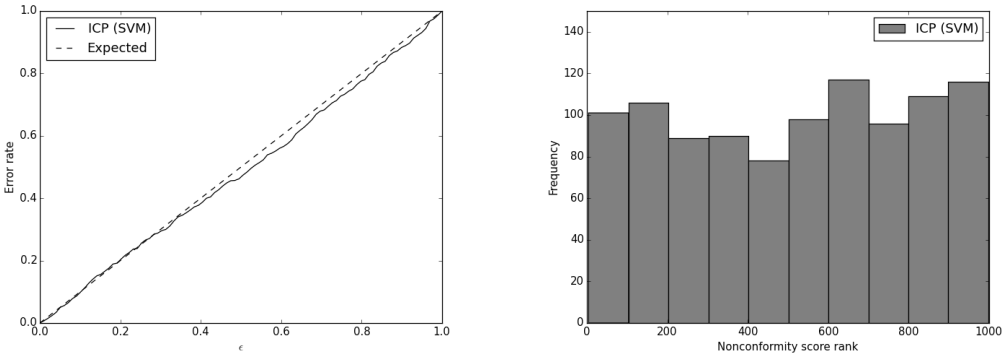
Figure 1: Distributions of p-values and nonconformity scores for 1000 test patterns from the spambase dataset for an inductive conformal predictor. The underlying model is a support vector machine, trained to produce class probability estimates, and the corresponding inductive conformal predictor is calibrated on 999 calibration examples using $f(\zeta, z_i) = 1 - \hat{P}_h(y_i \mid \boldsymbol{x_i})$.

The second criterion ensures that a conformal predictor is *automatically* well-calibrated (i.e., valid). If the property of being well-calibrated is independent on the choice of f, then the validity of the predictions made by a conformal predictor is dependent only on the assumption of the sequence being exchangeable (Vovk et al., 2006).

In the following sections, we will show that aggregated conformal predictors (including cross-conformal predictors and bootstrap conformal predictors) can indeed fulfill criterion **I**, in that they may be well-calibrated, but do not fulfill criterion **II**. We also provide a revised definition of aggregated conformal predictors, that shows an approximate validity given certain constraints.

3.1. Cross-Conformal Predictors and Bootstrap Conformal Predictors

Let $H^* = \{h_1, \ldots, h_k\}$ be the underlying models generated through the cross-conformal prediction procedure described in Section 2.2, and let $Z^* = \{Z_1, \ldots, Z_k\}$ be the calibration sets corresponding to each of these models. If we choose any pair $(h_l \in H^*, Z_l \in Z^*)$, we can define a simple inductive conformal predictor, using h_l as the underlying model, Z_l as the calibration set and Equation (9) to compute the p-values. We know from previous



(a) Errors rate of an inductive conformal predictor (b) Nonconformity score rank distribution

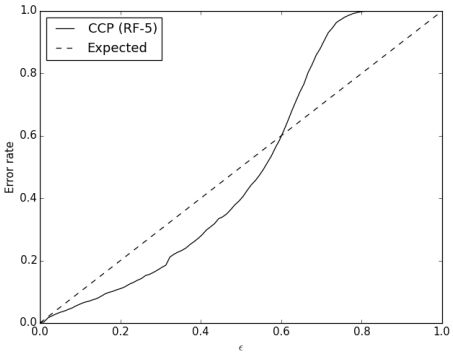
Figure 2: Calibration plot (empirical error rate) of an inductive conformal predictor on the spambase dataset, and distribution of ranks of the test patterns' nonconformity scores. The same type of conformal predictor was used as in Figure 1.

work that this inductive conformal predictor is valid (i.e., automatically well-calibrated), and fulfills both criteria regarding p-values given in the previous section (Vovk, 2013). Figure 2(a) shows the well-calibrated nature of such an inductive conformal predictor; for any value of ϵ, the observed error rate (rejection rate of true class labels) is very close to ϵ. Figure 2(b) shows the distribution of ranks of the test patterns' nonconformity scores, when testing for their correct label (a rank of r denotes that $r-1$ calibration examples had a larger nonconformity score than the test pattern, i.e., the rank effectively corresponds to the numerator of the p-value equation).

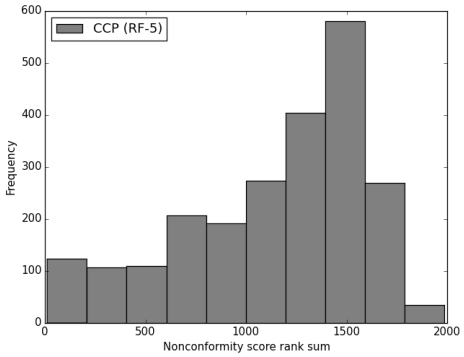
We now move to a (partial) cross-conformal predictor, by selecting two pairs of predictive models and calibration sets, $(h_l \in H^*, Z_l \in Z^*)$ and $(h_m \in H^*, Z_m \in Z^*)$, where $m \neq l$, and use Equation (12) to compute the p-values. Since any of the two pairs, (h_l, Z_l) or (h_m, Z_m), can be used to construct an inductive conformal predictor, we know that both of them will produce uniformly distributed ranks of the test patterns' nonconformity scores, as shown in Figure 2(b). Let $r^l_{n+1} \in \mathbb{Z}$ and $r^m_{n+1} \in \mathbb{Z}$ denote the ranks produced by each pair (we will denote these pairs as *ICP components* of the cross-conformal predictor), respectively, for the test pattern z_{n+1}. For a cross-conformal predictor to be well-calibrated, it is necessary that all sums $2 \leq \left(r^l_{n+1} + r^m_{n+1}\right) \in \mathbb{Z} \leq l + m + 2$ are equiprobable, unconditional on z_{n+1}—in Equation (12), we are effectively summing the ranks of the individual ICP components of the cross-conformal predictor in order to compute the p-value—only then is $p^{y_{n+1}}_{n+1}$ distributed according to $U[0, 1]$.

Since we wish for $r^* = r^l + r^m$ to be uniformly distributed, we are required to put constraints on the joint distribution of $\left(r^l, r^m\right)$. If we allow r^l and r^m to be distributed uniformly on the rectangular surface defined by l and m, i.e., the two ranks obtained from the ICP components are independent of each other, their sum r^* is no longer distributed uniformly, but instead distributed according to the Irwin-Hall (uniform sum) distribution (Irwin, 1927; Hall, 1927). p-values are then distributed according to the unimodal Bates

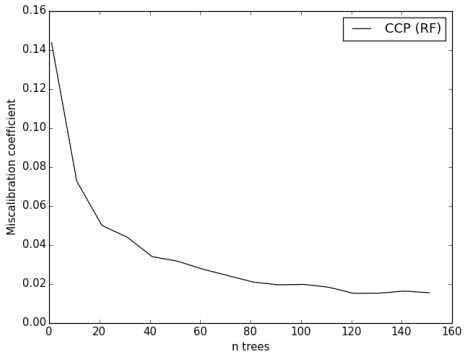
distribution (Bates et al., 1955) rather than $U[0, 1]$, such that p-values closer to the mean (i.e., 0.5) are more likely than extreme values (i.e., p-values closer to 0 or 1). Including more inductive conformal predictor components, by combining several pairs (h_*, Z_*), further increases this effect, as the variance of the Bates distribution decreases. This leads to a conformal predictor that is conservative for low values of ϵ (since small p-values are overly rare) and invalid for large values of ϵ (since large p-values are also rare).



(a) Cross-conformal predictor errors (random forest of 5 trees)



(b) Nonconformity score rank sum distribution (random forest of 5 trees)

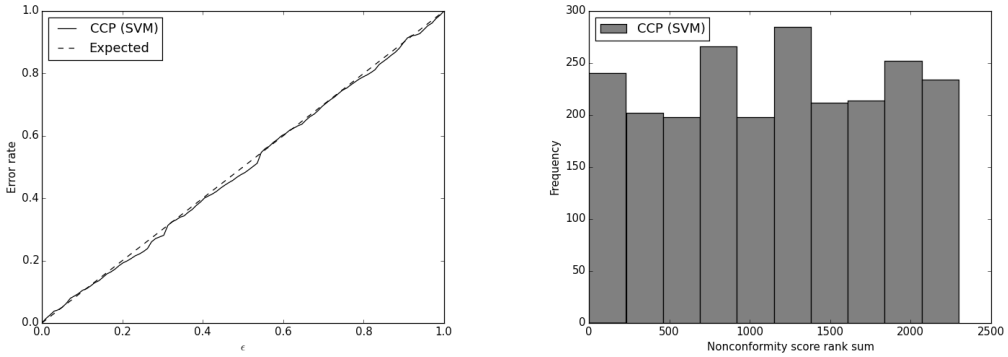


(c) Miscalibration rate depending on forest size

Figure 3: Calibration plot (empirical error rate) of a cross-conformal predictor ($k = 10$) on the spambase dataset, and distribution of rank sums of the test patterns' nonconformity scores. A random forest using 5 trees was used as the underlying model.

Figure 3 shows a poorly calibrated cross-conformal predictor ($k = 10$), where the underlying model in each fold is a weak random forest (containing only 5 trees). Since the underlying models are fairly unstable (i.e., highly variant depending on their particular training data), the sum of their ranks is far from uniformly distributed as shown in Figure 3(b). The calibration plot shown in Figure 3(a) illustrates the expected behaviour, with the cross-conformal predictor being conservative at low values of ϵ and invalid at large values of

ϵ. Figure 3(c) shows how the miscalibration rate—area between error curve and expected error rate (diagonal line) in Figure 3(a)—reduces with an increasing forest size; adding additional ensemble members to the random forest models reduces their variance, and the 10 random forest models (over the 10 folds) become more similar each other, resulting in the nonconformity rank sums approaching a uniform distribution.



(a) Cross-conformal predictor errors (svm) (b) Nonconformity score rank sum distribution (svm)

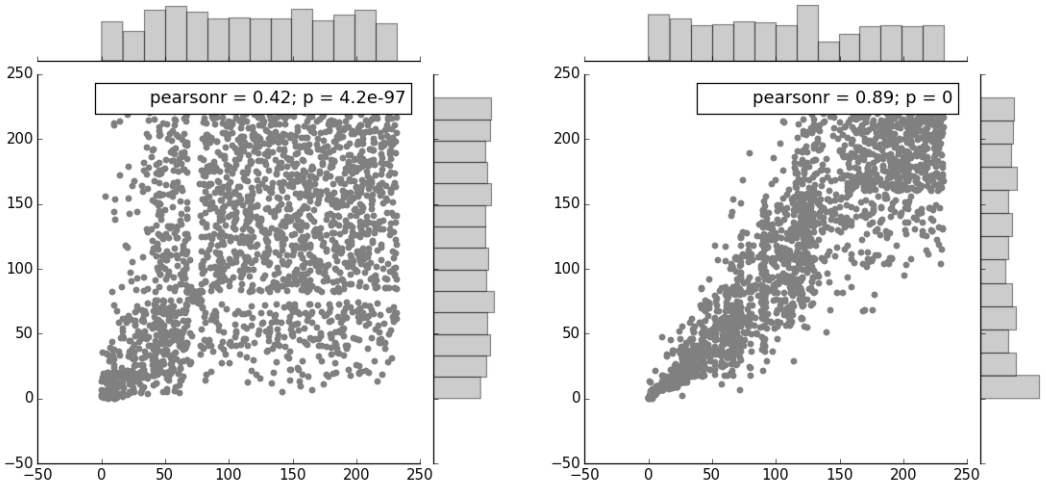
Figure 4: Calibration plot (empirical error rate) of a cross-conformal predictor ($k = 10$) on the spambase dataset, and distribution of rank sums of the test patterns' nonconformity scores. A support vector machine was used as the underlying model.

In Figure 4, a well-calibrated cross-conformal predictor ($k = 10$) is displayed, using support vector machines as the underlying models. With this setup, the rank sums, Figure 4(b), appear uniformly distributed. The error rates, Figure 4(a), are close or equal to ϵ over the entire range $\epsilon \in (0, 1)$, similar to the results obtained by Vovk (2013) using MART to construct the underlying models.

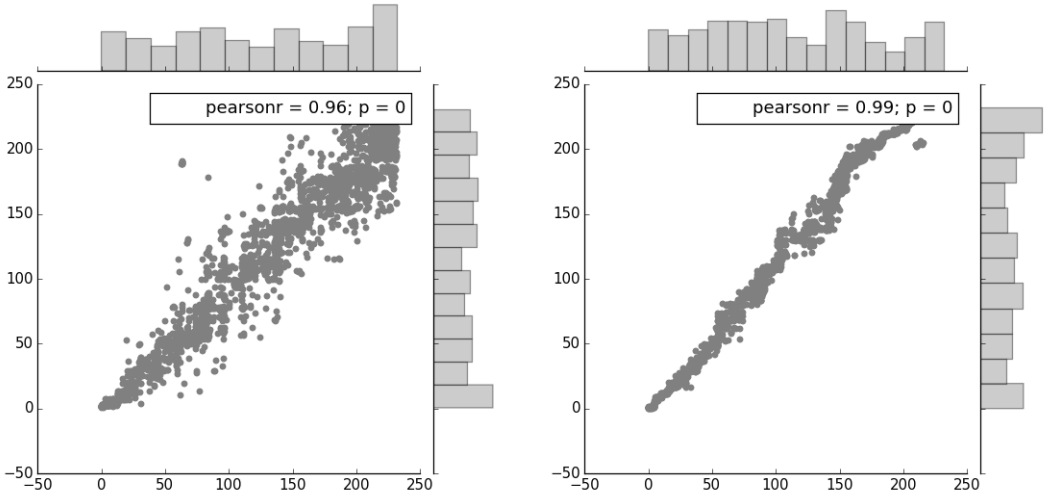
Figure 5 shows the distribution of nonconformity ranks and nonconformity rank sums from of pairs ICP components, i.e., r^l_{n+1} and r^m_{n+1}, from cross-conformal predictors ($k = 10$) created using various underlying models: random forests with 5, 100 and 500 trees, as well as a support vector machine. It is clear that: (1) regardless of the stability of the underlying models, a single ICP component provides uniformly distributed nonconformity ranks (top and rightmost histogram in each plot); and (2) the stability of the underlying models has a large impact on the joint distribution of $\left(r^l, r^m\right)$, with the most unstable underlying model (random forest using 5 trees) shows a near-uniform joint distribution of nonconformity ranks (middle scatter plots), resulting in an approximate Irwin-Hall distribution of nonconformity rank sums.

3.2. Aggregated Conformal Predictors

In Section 2.4, we provided the general definition of aggregated conformal predictors originally given by Carlsson et al. (2014), arguing that such aggregate models (including CCP



(*a*) Nonconformity ranks (rf, 5 trees) (*b*) Nonconformity ranks (rf, 100 trees)



(*c*) Nonconformity ranks (rf, 500 trees) (*d*) Nonconformity ranks (svm)

Figure 5: Distribution of ranks from two inductive conformal predictor components of a cross-conformal predictor ($k = 10$). In each plot, the top and right-side histograms show the rank distributions of two individual components, r^l and r^m, while the middle scatter plot shows the joint distribution of $\left(r^l, r^m\right)$.

and BCP) might provide valid aggregate p-values given a certain condition: that the calibration sets are consistently resampled with respect to a particular nonconformity score α_t. We noted also that, while the definition of consistent resampling is clear, how to obtain a consistent resampling is not. Finally, we argued—and showed—in Section 3.1 that the sought-after well-calibrated nature of conformal predictors is not automatically guaranteed for aggregate models; we must note also that the results in Section 3.1 are in accordance with the empirical results provided by Carlsson et al. (2014), where the ACP models were shown to be notably conservative for significance levels $\epsilon \leq 0.4$.

To shed some light on the nature of consistent resampling, we restate the condition under which aggregate conformal predictors are valid, using some additional definitions.

Definition 4 (Approximately ranking invariant) *Let Z^n be a sequence of examples drawn from the problem space $Z \sim \mathbb{P}$ and let $Z^m \subseteq Z^n$ be an exchangeable resampling of Z^n. Let H be a learning algorithm, and let f_n and f_m be nonconformity measures on the form*

$$f_s\left(Z^s, (\boldsymbol{x}_i, y_i,)\right) = \Delta\left(h^{-s}\left(\boldsymbol{x}_i\right), y_i\right), \tag{21}$$

where $h^{-s} = H\left(Z^{-s}\right)$, i.e., a predictive model trained using a proper training set $Z^{-s} \subset Z$ such that $Z^{-s} \cap Z^s = \emptyset$. Let r^n_{n+1} and r^m_{n+1} be the ranks produced by f_n and f_m for a test pattern $(\boldsymbol{x}_{n+1}, y_{n+1})$, using Z^n and Z^m as calibration sets respectively. f_s is approximately ranking invariant if, for any such f_n and f_m,

$$\bar{r}^m_{n+1} = \frac{r^m_{n+1}}{m+1} \approx \frac{r^n_{n+1}}{n+1} = \bar{r}^n_{n+1},$$

for finite $m \leq n$.

Definition 5 (Consistent mapping) *Let Z^n be a sequence drawn from the problem space $Z \sim \mathbb{P}$, let $Z^m \subseteq Z^n$ be an exchangeable resampling of Z^n, and let f be a nonconformity measure. f is a consistent mapping of Z if Z^m is a consistent resampling of Z^n with respect to $\bar{r}^n_{n+1}$.*

Remark 6 *Approximately ranking invariant nonconformity measures and consistent mappings are not interchangeable. We can think of rankings that appear similar in the finite case, but do not converge asymptotically. Similarly, we can think of rankings that start off dissimilar in the finite case but eventually converge in the limit.*

Proposition 7 *Aggregated conformal predictors are approximately valid when f is an approximately ranking invariant consistent mapping of Z.*

Proof *Let $Z_1, \ldots, Z_k$ be calibration sets exchangeably resampled from Z^n such that $\forall l \in (1, k) : Z_l \subset Z^n$, and let $f_1, \ldots, f_k$ be approximately ranking invariant consistent mappings constructed using the complementary proper training sets $Z_{-1}, \ldots, Z_{-k}$ where $\forall l \in (1, k) : Z_{-l} = Z^n \setminus Z_l$. Each mapping f_l consists of an underlying model h_l and a calibration set Z_l. Define an aggregate conformal predictor using the pairs $\{f_1 = (h_1, Z_1), \ldots, f_k = (h_k, Z_k)\}$ as the ICP components.*

Let Z_l and Z_m be two distinct ICP components. Since Z_l and Z_m are, by definition, exchangeably resampled from Z^n, f_l and f_m are valid ICPs that output p_{n+1}^{y+1}-values distributed according to $U[0,1]$; by extension, f_l and f_m output valid ranks r_{n+1} for the true class y_{n+1} distributed according to $U[0,|Z_l|+1]$ and $U[0,|Z_m|+1]$ respectively. f_l and f_m are both approximately ranking invariant and consistent mappings with respect to $\bar{r}_{n+1}$, i.e., $\bar{r}_{n+1}^l \approx \bar{r}_{n+1}^m \approx \bar{r}_{n+1}$ for finite $l, m \leq n$ and $\lim_{l,m,n\to\infty} \bar{r}_{n+1}^l = \bar{r}_{n+1}^m = \bar{r}_{n+1}$. Since $\bar{r}_{n+1} = p_{n+1}^{y+1}$, $f_l \cup f_m$ represents an asymptotically exact aggregate conformal predictor that approximates a valid conformal predictor for finite calibration sets. ■

Remark 8 *We are unable to provide a formal definition of approximate validity. From Definition 4, we ask that the nonconformity measures $f_1, \ldots, f_k$ produce similar rankings of the test pattern; since the condition is stated loosely (the ranks produced are approximately equal), we can only state the conclusion loosely in the finite case: the error rate of the aggregated conformal predictor is approximately equal to that of an inductive conformal predictor, i.e., it is close to ϵ.*

We note that we could just as well state in Definition 5 that a consistent resampling is required with respect to p_{n+1}^{n+1}, however, we wish to make explicit the restrictions that are put on f, and by extension, h. Given a test object $\boldsymbol{x}_{n+1}$, an exchangeably resampled calibration set Z_l, and a predictive model h_l trained on Z_{-l}, the performance of h_l must be essentially invariant on $\boldsymbol{x}_{n+1}$ relative to Z_l, regardless of the specific composition of Z_{-l}. That is, we require that the underlying learning algorithm is *stable* in the sense of Breiman (1996), i.e., that small changes in the training set must not cause large changes in the resulting model. This is very much in-line with remarks made earlier by Vovk (2015, Appendix A), who notes that the validity of leave-one-out conformal predictors (n-fold cross-conformal predictors) is dependent on the underlying models, such that the resulting aggregated conformal predictor is invalid if the n-fold nonconformity function is not transitive. Here, we have shown that this requirement is not unique to leave-one-out conformal predictors, but applies to aggregated conformal predictors in general. As also noted by (Vovk, 2015, Appendix A), validity is violated in a "non-interesting way", in that the resulting aggregated conformal predictor is invalid in a conservative manner for low values of ϵ, i.e., the empirical error rate is deflated rather than inflated. This means, on the one hand, that we can utilize aggregated conformal predictors without needing to worry about an exaggerated empirical error rate. On the other hand, conservatively valid conformal predictors are less useful to us than exactly valid conformal predictors, since we are not able to leverage the excess confidence; if we provide our conformal predictor with a significance value $\epsilon = 0.05$, we should still act as though 5% of all predictions are incorrect, even though this might not be the case in reality. For any conservative predictor we could arbitrarily reduce the size of output predictions until the error rate is exactly ϵ, hence, the predictions made by an aggregated conformal predictor are—if our nonconformity measure does not fulfill the criteria given in Proposition 7—by definition, unnecessarily large.

4. Conclusions and Future Work

In this paper, we have provided a thorough investigation into the validity of aggregated conformal predictors, considering the definitions of cross-conformal predictors and bootstrap conformal predictors provided by Vovk (2015) as well as the generalized definition provided by Carlsson et al. (2014). We conclude that the validity of any aggregate conformal predictor is conditional on the nonconformity measure, in particular its ability to consistently rank individual objects amongst a group of objects. If the nonconformity measure does not possess this characteristic, the resulting aggregated conformal predictor is only conservatively valid for interesting (low) values of ϵ, i.e., the empirical error rate is lower than the expectation. While this is beneficial from a safety standpoint, it also means that the predictions output by an aggregated conformal predictor may be unnecessarily large.

While the definitions provided in this paper provide some tools for reasoning about the validity of aggregated conformal predictors, they do not provide sufficient practical guidance. We have stated that the underlying model should be *stable*, as defined by Breiman (1996), but do not quantitatively investigate the relationship between instability and invalidity, nor have we assessed the effects of aggregating unstable nonconformity measures with respect to efficiency. In light of this, we propose that future work addresses the question of how to choose a suitable nonconformity measure, as well as investigates the magnitude of the negative effect on efficiency if an unsuitable nonconformity measure is selected.

We also propose that the aggregated conformal prediction scheme be evaluated in comparison to other methods of combining multiple underlying models, e.g., the provably valid bootstrap calibration procedure described by Boström et al. (2017) or other methods of combining p-values (some of which are addressed briefly in Appendix A).

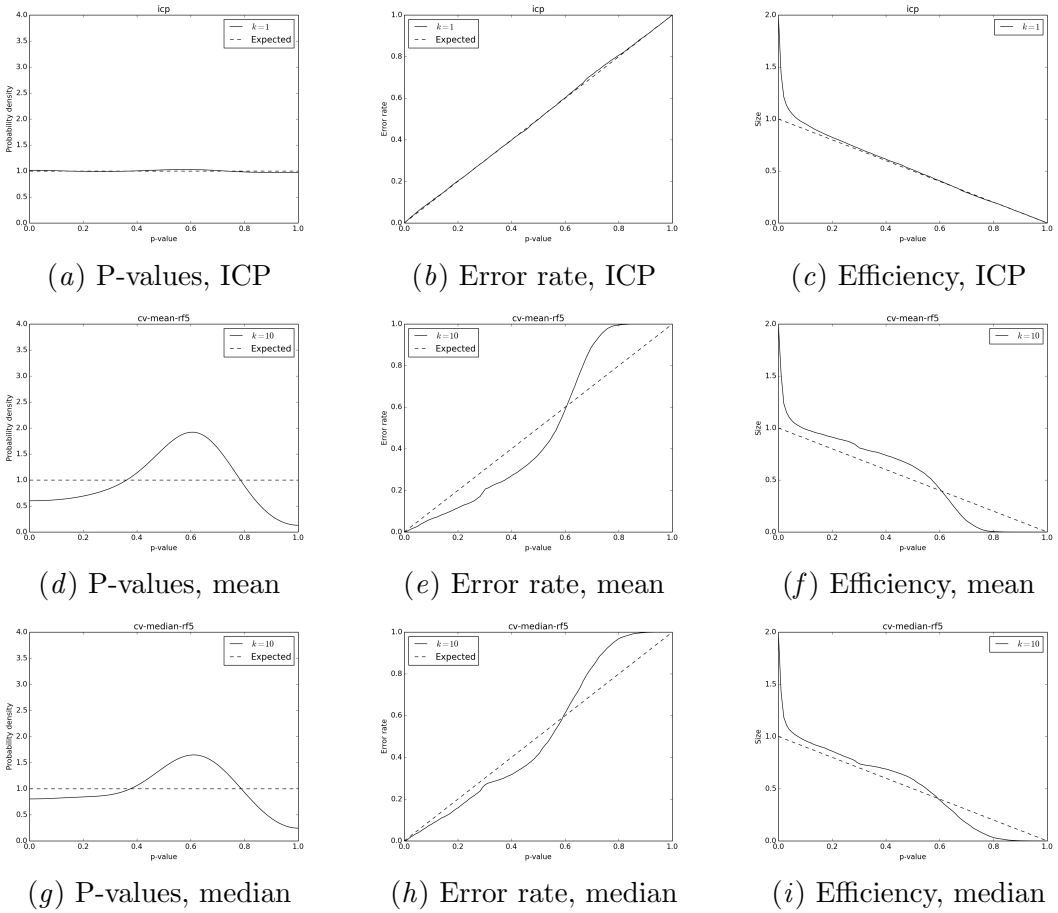
Acknowledgments

This work was supported by the Swedish Knowledge Foundation through the project Data Analytics for Research and Development (20150185). The research at Swetox (UN) was supported by Stockholm County Council, Knut & Alice Wallenberg Foundation, and Swedish Research Council FORMAS.

Appendix A. Alternative Methods of combining p-values

In this work, we have shown that aggregated conformal predictors are troublesome in that they are not valid (or, possibly, efficient) in general, but that we must put some constraints on the underlying model and the nonconformity measure we construct from it. The issues we see with aggregated conformal predictors stem from the fact that we are averaging p-values that show a varying degree of interdependence. It is thus natural to wonder whether our aggregated models could fare better if we, instead of combining the p-values through averaging, utilize some other aggregation procedure.

Figure 6 shows three variants of conformal predictors applied to the spambase data set. Figures 6(a) to 6(c) shows the distribution of p-values for the test set (for correct labels only), the empirical error rate and the efficiency of a simple ICP. Figures 6(d) to 6(f) show the analogous results from a k-folded aggregated conformal predictor, and Figures 6(g)



(*a*) P-values, ICP (*b*) Error rate, ICP (*c*) Efficiency, ICP

(*d*) P-values, mean (*e*) Error rate, mean (*f*) Efficiency, mean

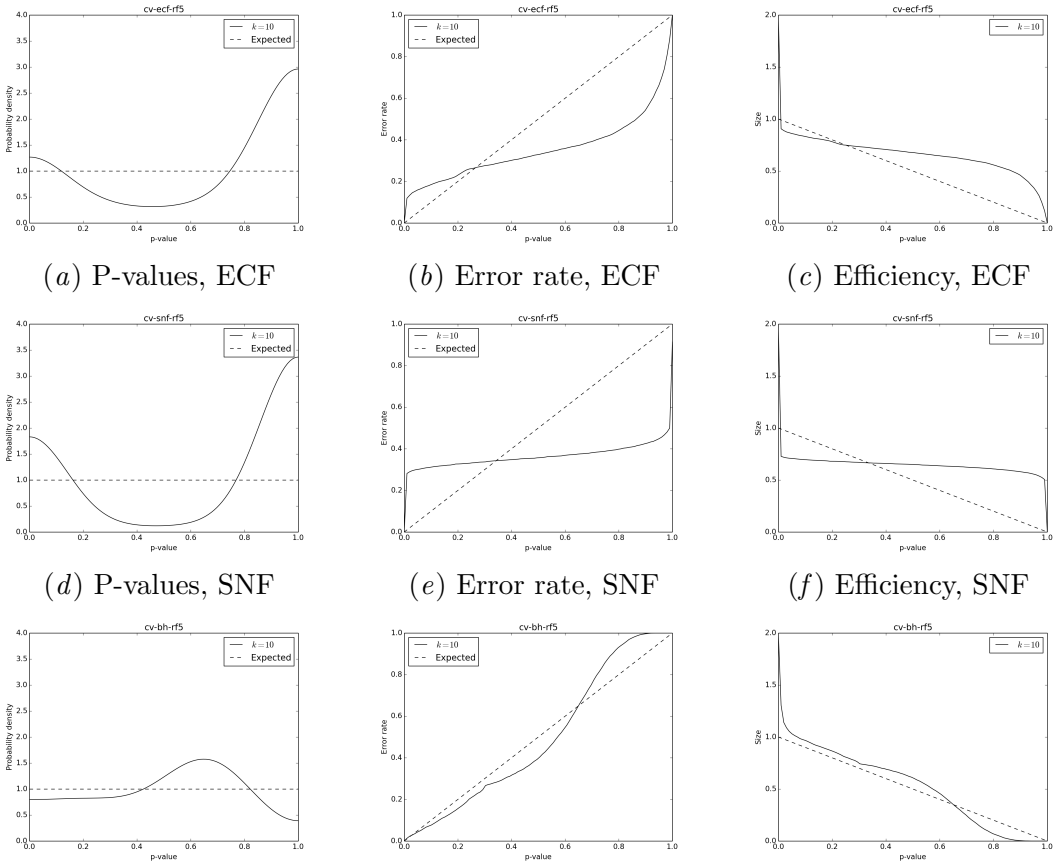
(*g*) P-values, median (*h*) Error rate, median (*i*) Efficiency, median

Figure 6: p-value distributions (of correct class), error rates and efficiency of various conformal predictors on the spambase data set. The conformal predictors are: an inductive conformal predictor (top), an aggregated conformal predictor (middle), a set of inductive conformal predictors whose p-values are aggregated by their median. Underlying models are random forests consisting of 5 trees.

to 6(i) show results from a set of k-folded conformal predictors outputting the median p-value rather than the mean. Here, we have used an underlying model previously identified as problematic with regard to aggregated conformal prediction: a random forest model consisting of only 5 trees, i.e., a model that whose decision function varies substantially based on the particular examples that are included in the training data. For the aggregate models, k was set to 10.

As noted previously, the distribution of p-values obtained from an aggregated conformal predictor, shown in Figure 6(d), is unimodal rather than uniform when the underlying model is unstable, which results in the sigmoidal error curve in Figure 6(e). Efficiency, shown in Figure 6(f) is clearly hampered in comparison to that of the ICP, shown in Figure 6(c). Note that the dashed lines in these last plots are provided simply to enable visual

comparison. Although combining p-values using their median rather than their mean, as shown in Figures 6(g) to 6(i), illustrates a similar behaviour, the negative effects on validity and efficiency appear lessened at all significance levels, suggesting that taking the median p-value might be a more suitable approach for constructing aggregated conformal predictors.



(a) P-values, ECF (b) Error rate, ECF (c) Efficiency, ECF

(d) P-values, SNF (e) Error rate, SNF (f) Efficiency, SNF

(g) P-values, BH (h) Error rate, BH (i) Efficiency, BH

Figure 7: p-value distributions (of correct class), error rates and efficiency of inductive conformal predictor sets whose p-values are combined using various methods. Results are taken from the spambase data set. The aggregation methods are: extended chi-square function (ECF, top), simple normal form (SNF, middle) and the Benjamini-Hochberg procedure (BH, bottom). Underlying models are random forests consisting of 5 trees.

Balasubramanian et al. (2015) propose several methods of constructing ensemble models from multiple conformal predictors, amongst them two procedures for combining the p-values obtained from multiple sources. The first, *extended chi-square function* (ECF) is

based on the work of Fisher (1948), defining the aggregated p-value as

$$p^* \sum_{i=0}^{k-1} \frac{(-\ln p^*)^i}{i!}, \tag{22}$$

where k is the number of p-values considered, and p^* is their product. A similar approach has also been described by Vovk (2015, Appendix C). The results of ECF are shown in Figures 7(a) to 7(c), where it is clear that the procedure is invalid for low values of ϵ.

The second approach proposed by Balasubramanian et al. (2015) is the *standard normal form* (SNF), where p-values obtained from the conformal predictors are combined by computing, for each p-value, the inverse of the normal CDF, $q_i = F^{-1}(p_i)$, taking the sum $q^* = \sum q_i$, and finally computing the aggregated p-value, again using the normal CDF, as $p = F(q_i)$. Similarly to ECF, this approach also shows, in Figures 7(d) to 7(f), invalid results for the spambase data set at low significance levels.

Finally, in figures 7(g) to 7(i), we evaluate a p-value combination method based on the Benjamini-Hochberg procedure for false discovery rate correction (Benjamini and Hochberg, 1995). Here, the aggregated p-value is defined as

$$p = \min \sum_{i=1}^{k} p_i \frac{k}{i}, \tag{23}$$

where $p_1, \ldots, p_k$ are sorted in ascending order. While this approach appears empirically sound when applied to the spambase data set, it does not fare better than combining the p-values through their median in terms of conservativeness or efficiency.

References

Vineeth N Balasubramanian, Shayok Chakraborty, and Sethuraman Panchanathan. Conformal predictions for information fusion. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):45–65, 2015.

Grace E Bates et al. Joint distributions of time intervals for the occurrence of successive accidents in a generalized polya scheme. *The Annals of Mathematical Statistics*, 26(4): 705–720, 1955.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Siddhartha Bhattacharyya. Confidence in predictions from random tree ensembles. *Knowledge and information systems*, 35(2):391–410, 2013.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, pages 1–20, 2017.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *COLT*, pages 605–622, 2014.

Lars Carlsson, Martin Eklund, and Ulf Norinder. Aggregated conformal prediction. In *Artificial Intelligence Applications and Innovations*, pages 231–240. Springer, 2014.

Dmitry Devetyarov and Ilia Nouretdinov. Prediction with confidence based on a random forest classifier. In *Artificial Intelligence Applications and Innovations*, pages 37–44. Springer, 2010.

Ronald A Fisher. Combining independent tests of significance. *American Statistician*, 2(5): 30, 1948.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning the second computer journal lecture. *The Computer Journal*, 50(2):151–163, 2007.

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 148–155. Morgan Kaufmann Publishers Inc., 1998.

Philip Hall. The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, pages 240–245, 1927.

Joseph Oscar Irwin. On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to pearson's type ii. *Biometrika*, pages 225–239, 1927.

Ulf Johansson, Henrik Boström, and Tuve Löfström. Conformal prediction using decision trees. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 330–339. IEEE, 2013.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176, 2014.

Ulf Johansson, Cecilia Sönströd, and Henrik Linusson. Efficient conformal regressors using bagged neural nets. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8. IEEE, 2015.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *Artificial Intelligence Applications and Innovations*, pages 261–270. Springer, 2014.

Tuve Löfström, Ulf Johansson, and Henrik Boström. Effective utilization of data in inductive conformal prediction using ensembles of neural networks. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–8. IEEE, 2013.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. *Tools in artificial intelligence*, 18(315-330):2, 2008.

Harris Papadopoulos. Cross-conformal prediction with ridge regression. In *Statistical Learning and Data Sciences*, pages 260–270. Springer, 2015.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40(1): 815–840, 2011.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, volume 2, pages 722–726, 1999.

Paolo Toccaceli, Ilia Nouretdinov, and Alexander Gammerman. Conformal predictors for compound activity prediction. In *Symposium on Conformal and Probabilistic Prediction with Applications*, pages 51–66. Springer, 2016.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine learning*, 92(2-3):349–376, 2013.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Verlag, DE, 2006.