JMLR: Workshop and Conference Proceedings 25:475–490, 2012 Asian Conference on Machine Learning

Conditional validity of inductive conformal predictors

Vladimir Vovk V.VOVK@RHUL.AC.UK
Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

Editors: Steven C. H. Hoi and Wray Buntine

Abstract

Conformal predictors are set predictors that are automatically valid in the sense of having coverage probability equal to or exceeding a given confidence level. Inductive conformal predictors are a computationally efficient version of conformal predictors satisfying the same property of validity. However, inductive conformal predictors have been only known to control unconditional coverage probability. This paper explores various versions of conditional validity and various ways to achieve them using inductive conformal predictors and their modifications.

Keywords: Inductive conformal predictors, conditional validity, batch mode of learning, boosting, MART, spam detection

1. Introduction

This paper continues study of the method of conformal prediction (Vovk et al. 2005, Chapter 2). An advantage of the method is that its predictions (which are set rather than point predictions) automatically satisfy a finite-sample property of validity. Its disadvantage is its relative computational inefficiency in many situations. A modification of conformal predictors, called inductive conformal predictors (Vovk et al. 2005, Section 4.1) aims at improving on the computational efficiency of conformal predictors.

Most of the literature on conformal prediction studies the behavior of set predictors in the online mode of prediction, perhaps because the property of validity can be stated in an especially strong form in the on-line mode (Vovk et al. 2005, Proposition 2.3). The online mode, however, is much less popular in applications of machine learning than the batch mode of prediction. This paper follows the recent papers by Lei et al. (2011), Lei and Wasserman (2012), and Lei et al. (2012) studying properties of conformal prediction in the batch mode; we, however, concentrate on inductive conformal prediction (also considered in Lei et al. 2012). Its full version is published as Vovk (2012).

We will usually be making the *assumption of randomness*, which is standard in machine learning and nonparametric statistics: the available data is a sequence of *examples* generated independently from the same probability distribution P. (In some cases we will make the weaker assumption of exchangeability; for some of our results even weaker assumptions, such as conditional randomness or exchangeability, would have been sufficient.) Each example consists of two components: an *object* and a *label*. We are given a *training set* of examples and a new object, and our goal is to predict the label of the new object. (If we have a whole

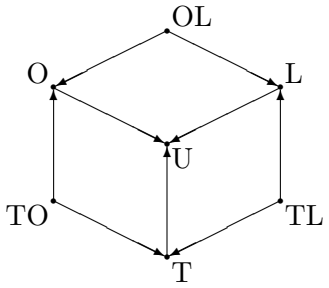

Figure 1: Eight notions of conditional validity. The visible vertices of the cube are U (unconditional), T (training conditional), O (object conditional), L (label conditional), OL (example conditional), TL (training and label conditional), TO (training and object conditional). The invisible vertex is TOL (and corresponds to conditioning on everything).

test set of new objects, we can apply the procedure for predicting one new object to each of the objects in the test set.)

The two desiderata for inductive conformal predictors are their validity and efficiency: validity requires that the coverage probability of the prediction sets should be at least equal to a preset confidence level, and efficiency requires that the prediction sets should be as small as possible. However, there is a wide variety of notions of validity, since the "coverage probability" is, in general, conditional probability. The simplest case is where we condition on the trivial σ-algebra, i.e., the probability is in fact unconditional probability, but several other notions of conditional validity are depicted in Figure 1, where T refers to conditioning on the training set, O to conditioning on the test object, and L to conditioning on the test label. The arrows in Figure 1 lead from stronger to weaker notions of conditional validity; U is the sink and TOL is the source (the latter is not shown).

Inductive conformal predictors will be defined in Section 2. They are automatically valid, in the sense of unconditional validity. It should be said that, in general, the unconditional error probability is easier to deal with than conditional error probabilities; e.g., the standard statistical methods of cross-validation and bootstrap provide decent estimates of the unconditional error probability but poor estimates for the training conditional error probability: see Hastie et al. (2009), Section 7.12.

In Section 3 we explore training conditional validity of inductive conformal predictors. Our simple results (Propositions 2a and 2b) are of the PAC type, involving two parameters: the target training conditional coverage probability $1 - \epsilon$ and the probability $1 - \delta$ with which $1 - \epsilon$ is attained. They show that inductive conformal predictors achieve training conditional validity automatically (whereas for other notions of conditional validity the method has to be modified). We give self-contained proofs of Propositions 2a and 2b, but Appendix A of Vovk (2012) explains how they can be deduced from classical results about tolerance regions.

In the following section, Section 4, we introduce a conditional version of inductive conformal predictors and explain, in particular, how it achieves label conditional validity. Label conditional validity is important as it allows the learner to control the set-prediction analogues of false positive and false negative rates. Section 5 is about object conditional validity

and its main result (a version of a lemma in Lei and Wasserman 2012) is negative: precise object conditional validity cannot be achieved in a useful way unless the test object has a positive probability. Whereas precise object conditional validity is usually not achievable, we should aim for approximate and asymptotic object conditional validity when given enough data (cf. Lei and Wasserman 2012).

Section 6 reports on the results of empirical studies for the standard `Spambase` data set (see, e.g., Hastie et al. 2009, Chapter 1, Example 1, and Section 9.1.2). Section 7 discusses close connections between an important class of ICPs and ROC curves. Section 8 concludes.

2. Inductive conformal predictors

The example space will be denoted $\mathbf{Z}$; it is the Cartesian product $\mathbf{X} \times \mathbf{Y}$ of two measurable spaces, the object space and the label space. In other words, each example $z \in \mathbf{Z}$ consists of two components: $z = (x, y)$, where $x \in \mathbf{X}$ is its object and $y \in \mathbf{Y}$ is its label. Two important special cases are the problem of *classification*, where $\mathbf{Y}$ is a finite set (equipped with the discrete σ-algebra), and the problem of *regression*, where $\mathbf{Y} = \mathbb{R}$.

Let $(z_1, \dots, z_l)$ be the training set, $z_i = (x_i, y_i) \in \mathbf{Z}$. We split it into two parts, the *proper training set* $(z_1, \dots, z_m)$ of size $m < l$ and the *calibration set* of size $l - m$. An *inductive conformity m-measure* is a measurable function $A : \mathbf{Z}^m \times \mathbf{Z} \to \mathbb{R}$; the idea behind the *conformity score* $A((z_1, \dots, z_m), z)$ is that it should measure how well z conforms to the proper training set. A standard choice is

$$A((z_1, \dots, z_m), (x, y)) := \Delta(y, f(x)), \tag{1}$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is a prediction rule found from $(z_1, \dots, z_m)$ as the training set and $\Delta : \mathbf{Y} \times \mathbf{Y}' \to \mathbb{R}$ is a measure of similarity between a label and a prediction. Allowing $\mathbf{Y}'$ to be different from $\mathbf{Y}$ (often $\mathbf{Y}' \supset \mathbf{Y}$) may be useful when the underlying prediction method gives additional information to the predicted label; e.g., the MART procedure used in Section 6 gives the logit of the predicted probability that the label is 1.

The *inductive conformal predictor* (ICP) corresponding to A is defined as the set predictor

$$\Gamma^\epsilon(z_1, \dots, z_l, x) := \{y \mid p^y > \epsilon\}, \tag{2}$$

where $\epsilon \in [0, 1]$ is the chosen *significance level* ($1 - \epsilon$ is known as the *confidence level*), the *p-values* p^y, $y \in \mathbf{Y}$, are defined by

$$p^y := \frac{|\{i = m + 1, \dots, l \mid \alpha_i \le \alpha^y\}| + 1}{l - m + 1}, \tag{3}$$

and

$$\alpha_i := A((z_1, \dots, z_m), z_i), \quad i = m + 1, \dots, l, \qquad \alpha^y := A((z_1, \dots, z_m), (x, y)) \tag{4}$$

are the conformity scores. Given the training set and a new object x the ICP predicts its label y; it *makes an error* if $y \notin \Gamma^\epsilon(z_1, \dots, z_l, x)$.

The random variables whose realizations are x_i, y_i, z_i, z will be denoted by the corresponding upper case letters (X_i, Y_i, Z_i, Z, respectively). The following proposition of validity is almost obvious.

Proposition 1 (Vovk et al., 2005, Proposition 4.1) *If random examples $Z_{m+1}, \dots, Z_l$, $Z_{l+1} = (X_{l+1}, Y_{l+1})$ are exchangeable (i.e., their distribution is invariant under permutations), the probability of error $Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1})$ does not exceed ϵ for any ϵ and any inductive conformal predictor Γ.*

In practice the probability of error is usually close to ϵ (as we will see in Section 6).

3. Training conditional validity

As discussed in Section 1, the property of validity of inductive conformal predictors is unconditional. The property of conditional validity can be formalized using a PAC-type 2-parameter definition. It will be convenient to represent the ICP (2) in a slightly different form downplaying the structure (x_i, y_i) of z_i. Define $\Gamma^\epsilon(z_1, \dots, z_l) := \{(x, y) \mid p^y > \epsilon\}$, where p^y is defined, as before, by (3) and (4) (therefore, p^y depends implicitly on x). Proposition 1 can be restated by saying that the probability of error $Z_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l)$ does not exceed ϵ provided $Z_1, \dots, Z_{l+1}$ are exchangeable.

We consider a canonical probability space in which $Z_i = (X_i, Y_i)$, $i = 1, \dots, l+1$, are i.i.d. random examples. A set predictor Γ (outputting a subset of $\mathbf{Z}$ given l examples and measurable in a suitable sense) is (ϵ, δ)-*valid* if, for any probability distribution P on $\mathbf{Z}$,

$$P^l \left(P(\Gamma(Z_1, \dots, Z_l)) \geq 1 - \epsilon \right) \geq 1 - \delta.$$

It is easy to see that ICPs satisfy this property for suitable ϵ and δ.

Proposition 2a *Suppose $\epsilon, \delta \in [0, 1]$,*

$$E \geq \epsilon + \sqrt{\frac{-\ln \delta}{2n}}, \tag{5}$$

where $n := l - m$ is the size of the calibration set, and Γ is an inductive conformal predictor. The set predictor Γ^ϵ is then (E, δ)-valid. Moreover, for any probability distribution P on $\mathbf{Z}$ and any proper training set $(z_1, \dots, z_m) \in \mathbf{Z}^m$,

$$P^n \left(P(\Gamma(z_1, \dots, z_m, Z_{m+1}, \dots, Z_l)) \geq 1 - \epsilon \right) \geq 1 - \delta.$$

This proposition gives the following recipe for constructing (ϵ, δ)-valid set predictors. The recipe only works if the training set is sufficiently large; in particular, its size l should significantly exceed $N := (-\ln \delta)/(2\epsilon^2)$. Choose an ICP Γ with the size n of the calibration set exceeding N. Then the set predictor $\Gamma^{\epsilon - \sqrt{(-\ln \delta)/(2n)}}$ will be (ϵ, δ)-valid.

Proof of Proposition 2a Let $E \in (\epsilon, 1)$ (not necessarily satisfying (5)). Fix the proper training set $(z_1, \dots, z_m)$. By (2) and (3), the set predictor Γ^ϵ makes an error, $z_{l+1} \notin \Gamma^\epsilon(z_1, \dots, z_l)$, if and only if the number of $i = m+1, \dots, l$ such that $\alpha_i \leq \alpha^y$ is at most $\lfloor \epsilon(n+1) - 1 \rfloor$; in other words, if and only if $\alpha^y < \alpha_{(k)}$, where $\alpha_{(k)}$ is the kth smallest α_i and $k := \lfloor \epsilon(n+1) - 1 \rfloor + 1$. Therefore, the P-probability of the complement of $\Gamma^\epsilon(z_1, \dots, z_l)$ is $P(A((z_1, \dots, z_m), Z) < \alpha_{(k)})$, where A is the inductive conformity m-measure. Set

$$\alpha^* := \inf\{\alpha \mid P(A((z_1, \dots, z_m), Z) < \alpha) > E\} \text{ and } \begin{cases} E' := P(A((z_1, \dots, z_m), Z) < \alpha^*) \\ E'' := P(A((z_1, \dots, z_m), Z) \leq \alpha^*). \end{cases}$$

The σ-additivity of measures implies that $E' \le E \le E''$, and $E' = E = E''$ unless α^* is an atom of $A((z_1, \dots, z_m), Z)$. Both when $E' = E$ and when $E' < E$, the probability of error will exceed E if an only if $\alpha_{(k)} > \alpha^*$. In other words, if only if we have at most $k-1$ of the α_i below or equal to α^*. The probability that at most $k - 1 = \lfloor \epsilon(n+1) - 1 \rfloor$ values of the α_i are below or equal to α^* equals $\mathbb{P}(B''_n \le \lfloor \epsilon(n+1) - 1 \rfloor) \le \mathbb{P}(B_n \le \lfloor \epsilon(n+1) - 1 \rfloor)$, where $B''_n \sim \mathrm{bin}_{n,E''}$, $B_n \sim \mathrm{bin}_{n,E}$, and $\mathrm{bin}_{n,p}$ stands for the binomial distribution with n trials and probability of success p. By Hoeffding's inequality (see, e.g., Vovk et al. 2005, p. 287), the probability of error will exceed E with probability at most

$$\mathbb{P}(B_n \le \lfloor \epsilon(n+1) - 1 \rfloor) \le \mathbb{P}(B_n \le \epsilon n) \le e^{-2(E-\epsilon)^2 n}. \tag{6}$$

Solving $e^{-2(E-\epsilon)^2 n} = \delta$ we obtain that Γ^ϵ is (E, δ)-valid whenever (5) is satisfied. ■

The inequality (5) in Proposition 2a is simple but somewhat crude as its derivation uses Hoeffding's inequality. The following proposition is the more precise version of Proposition 2a that stops short of that last step.

Proposition 2b *Let $\epsilon, \delta, E \in [0,1]$. If Γ is an inductive conformal predictor, the set predictor Γ^ϵ is (E, δ)-valid provided*

$$\delta \ge \mathrm{bin}_{n,E}\left(\lfloor \epsilon(n+1) - 1 \rfloor\right), \tag{7}$$

where $n := l - m$ is the size of the calibration set and $\mathrm{bin}_{n,E}$ is the cumulative binomial distribution function with n trials and probability of success E. If the random variable $A((z_1, \dots, z_m), Z)$ is continuous, Γ^ϵ is (E, δ)-valid if and only if (7) holds.

Proof See the left-most expression in (3) and remember that $E'' = E$ unless α^* is an atom of $A((z_1, \dots, z_m), Z)$. ■

4. Conditional inductive conformal predictors

The motivation behind conditional inductive conformal predictors is that ICPs do not always achieve the required probability ϵ of error $Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1})$ conditional on $(X_{l+1}, Y_{l+1}) \in E$ for important sets $E \subseteq \mathbf{Z}$. This is often undesirable. If, e.g., our set predictor is valid at the significance level 5% but makes an error with probability 10% for men and 0% for women, both men and women can be unhappy with calling 5% the probability of error. Moreover, in many problems we might want different significance levels for different regions of the example space: e.g., in the problem of spam detection (considered in Section 6) classifying spam as email usually makes much less harm than classifying email as spam.

An *inductive m-taxonomy* is a measurable function $K : \mathbf{Z}^m \times \mathbf{Z} \to \mathbf{K}$, where $\mathbf{K}$ is a measurable space. Usually the *category* $K((z_1, \dots, z_m), z)$ of an example z is a kind of classification of z, which may depend on the proper training set $(z_1, \dots, z_m)$.

The *conditional inductive conformal predictor* (conditional ICP) corresponding to K and an inductive conformity m-measure A is defined as the set predictor (2), where the p-values p^y are now defined by

$$p^y := \frac{|\{i = m+1, \dots, l \mid \kappa_i = \kappa^y \ \& \ \alpha_i \le \alpha^y\}| + 1}{|\{i = m+1, \dots, l \mid \kappa_i = \kappa^y\}| + 1}, \tag{8}$$

the categories κ are defined by

$$\kappa_i := K((z_1, \dots, z_m), z_i), \quad i = m+1, \dots, l, \qquad \kappa^y := K((z_1, \dots, z_m), (x, y)),$$

and the conformity scores α are defined as before by (4). A *label conditional ICP* is a conditional ICP with the inductive m-taxonomy $K(\cdot, (x, y)) := y$.

The following proposition is the conditional analogue of Proposition 1; in particular, it shows that in classification problems label conditional ICPs achieve label conditional validity.

Proposition 3 *If random examples $Z_{m+1}, \dots, Z_l, Z_{l+1} = (X_{l+1}, Y_{l+1})$ are exchangeable, the probability of error $Y_{l+1} \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X_{l+1})$ given the category $K((Z_1, \dots, Z_m), Z_{l+1})$ of Z_{l+1} does not exceed ϵ for any ϵ and any conditional inductive conformal predictor Γ corresponding to K.*

5. Object conditional validity

In this section we prove a negative result (a version of Lemma 1 in Lei and Wasserman 2012) which says that the requirement of precise object conditional validity cannot be satisfied in a non-trivial way for rich object spaces (such as $\mathbb{R}$). If P is a probability distribution on $\mathbf{Z}$, we let $P_{\mathbf{X}}$ stand for its marginal distribution on $\mathbf{X}$: $P_{\mathbf{X}}(A) := P(A \times \mathbf{Y})$. Let us say that a set predictor Γ *has* $1 - \epsilon$ *object conditional validity*, where $\epsilon \in (0, 1)$, if, for all probability distributions P on $\mathbf{Z}$ and $P_{\mathbf{X}}$-almost all $x \in \mathbf{X}$,

$$P^{l+1}\left(Y_{l+1} \in \Gamma(Z_1, \dots, Z_l, X_{l+1}) \mid X_{l+1} = x\right) \ge 1 - \epsilon. \tag{9}$$

The Lebesgue measure on $\mathbb{R}$ will be denoted Λ. If Q is a probability distribution, we say that a property F holds for Q-*almost all* elements of a set E if $Q(E \setminus F) = 0$; a Q-*non-atom* is an element x such that $Q(\{x\}) = 0$.

Proposition 4 *Suppose $\mathbf{X}$ is a separable metric space equipped with the Borel σ-algebra. Let $\epsilon \in (0, 1)$. Suppose that a set predictor Γ has $1 - \epsilon$ object conditional validity. In the case of regression, we have, for all P and for $P_{\mathbf{X}}$-almost all $P_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$,*

$$P^l\left(\Lambda(\Gamma(Z_1, \dots, Z_l, x)) = \infty\right) \ge 1 - \epsilon. \tag{10}$$

In the case of classification, we have, for all P, all $y \in \mathbf{Y}$, and $P_{\mathbf{X}}$-almost all $P_{\mathbf{X}}$-non-atoms x,

$$P^l\left(y \in \Gamma(Z_1, \dots, Z_l, x)\right) \ge 1 - \epsilon. \tag{11}$$

We are mainly interested in the case of a small ϵ (corresponding to high confidence), and in this case (10) implies that, in the case of regression, prediction intervals (i.e., the convex hulls of prediction sets) can be expected to be infinitely long unless the new object is an atom. In the case of classification, (11) says that each particular $y \in \mathbf{Y}$ is likely to be included in the prediction set, and so the prediction set is likely to be large. In particular, (11) implies that the expected size of the prediction set is a least $(1-\epsilon)\,|\mathbf{Y}|$.

Of course, the condition that x be a non-atom is essential: if $P_{\mathbf{X}}(\{x\}) > 0$, an inductive conformal predictor that ignores all examples with objects different from x will have $1-\epsilon$ object conditional validity and can give narrow predictions if the training set is big enough to contain many examples with x as their object.

Proof of Proposition 4 The proof will be based on the ideas of Lei and Wasserman (2012, the proof of Lemma 1).

Suppose (10) does not hold on a measurable set E of $P_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$ such that $P_{\mathbf{X}}(E) > 0$. Shrink E in such a way that $P_{\mathbf{X}}(E) > 0$ still holds but there exists $\delta > 0$ and $C > 0$ such that, for each $x \in E$,

$$P^l\left(\Lambda(\Gamma(Z_1,\dots,Z_l,x)) \le C\right) \ge \epsilon + \delta. \tag{12}$$

Let V be the total variation distance between probability measures, $V(P,Q) := \sup_A |P(A) - Q(A)|$; we then have

$$V(P^l, Q^l) \le \sqrt{2}\sqrt{1-(1-V(P,Q))^l}$$

(this follows from the connection of V with the Hellinger distance: see, e.g., Tsybakov 2010, Section 2.4). Shrink E further so that $P_{\mathbf{X}}(E) > 0$ still holds but

$$\sqrt{2}\sqrt{1-(1-P_{\mathbf{X}}(E))^l} \le \delta/2. \tag{13}$$

(This can be done under our assumption that $\mathbf{X}$ is a separable metric space: we can take the intersection of E and some neighbourhood of any element of $\mathbf{X}$ for which all such intersections have a positive $P_{\mathbf{X}}$-probability.) Define another probability distribution Q on $\mathbf{Z}$ by the requirements that $Q(A \times B) = P(A \times B)$ for all measurable $A \subseteq (\mathbf{X} \setminus E)$, $B \subseteq \mathbb{R}$ and $Q(A \times B) = P_{\mathbf{X}}(A) \times U(B)$ for all measurable $A \subseteq E$, $B \subseteq \mathbb{R}$, where U is the uniform probability distribution on the interval $[-DC, DC]$ and $D > 0$ will be chosen below. Since $V(P,Q) \le P_{\mathbf{X}}(E)$, we have $V(P^l, Q^l) \le \delta/2$; therefore, by (12),

$$Q^l\left(\Lambda(\Gamma(Z_1,\dots,Z_l,x)) \le C\right) \ge \epsilon + \delta/2$$

for each $x \in E$. The last inequality implies, by Fubini's theorem,

$$Q^{l+1}\left(\Lambda(\Gamma(Z_1,\dots,Z_l,X_{l+1})) \le C \;\&\; X_{l+1} \in E\right) \ge (\epsilon + \delta/2)\, Q_{\mathbf{X}}(E),$$

where $Q_{\mathbf{X}}(E) = P_{\mathbf{X}}(E) > 0$ is the marginal Q-probability of E. When $D = D(\delta Q_{\mathbf{X}}(E), C)$ is sufficiently large this in turn implies

$$Q^{l+1}\left(Y_{l+1} \notin \Gamma(Z_1,\dots,Z_l,X_{l+1}) \;\&\; X_{l+1} \in E\right) \ge (\epsilon + \delta/4)\, Q_{\mathbf{X}}(E).$$

However, the last inequality contradicts

$$\frac{Q^{l+1}\left(Y_{l+1} \notin \Gamma(Z_1, \dots, Z_l, X_{l+1}) \,\&\, X_{l+1} \in E\right)}{Q_{\mathbf{X}}(E)} \le \epsilon,$$

which follows from Γ having $1-\epsilon$ object conditional validity and the definition of conditional probability.

It remains to consider the case of classification. Suppose (11) does not hold on a measurable set E of $P_{\mathbf{X}}$-non-atoms $x \in \mathbf{X}$ such that $P_{\mathbf{X}}(E) > 0$. Shrink E in such a way that $P_{\mathbf{X}}(E) > 0$ still holds but there exists $\delta > 0$ such that, for each $x \in E$,

$$P^l\left(y \in \Gamma(Z_1, \dots, Z_l, x)\right) \le 1 - \epsilon - \delta.$$

Without loss of generality we further assume that (13) also holds. Define a probability distribution Q on $\mathbf{Z}$ by the requirements that $Q(A \times B) = P(A \times B)$ for all measurable $A \subseteq (\mathbf{X} \setminus E)$ and all $B \subseteq \mathbf{Y}$ and that $Q(A \times \{y\}) = P_{\mathbf{X}}(A)$ for all measurable $A \subseteq E$ (i.e., modify P setting the conditional distribution of Y given $X \in E$ to the unit mass concentrated at y). Then for each $x \in E$ we have

$$Q^l\left(y \in \Gamma(Z_1, \dots, Z_l, x)\right) \le 1 - \epsilon - \delta/2,$$

which implies

$$Q^{l+1}\left(Y_{l+1} \in \Gamma(Z_1, \dots, Z_l, X_{l+1}) \,\&\, X_{l+1} \in E\right) \le (1 - \epsilon - \delta/2)\, Q_{\mathbf{X}}(E).$$

The last inequality contradicts Γ having $1-\epsilon$ object conditional validity. ■

Proposition 4 can be extended to randomized set predictors Γ (in which case P^l and P^{l+1} in expressions such as (9) and (10) should be replaced by the probability distribution comprising both P and the internal coin tossing of Γ). This clarifies the provenance of ϵ in (10) and (11): ϵ cannot be replaced by a smaller constant since the set predictor predicting $\mathbf{Y}$ with probability $1-\epsilon$ and $\emptyset$ with probability ϵ has $1-\epsilon$ object conditional validity.

Proposition 4 does not prevent the existence of efficient set predictors that are conditionally valid in an asymptotic sense; indeed, the paper by Lei and Wasserman (2012) is devoted to constructing asymptotically efficient and asymptotically conditionally valid set predictors in the case of regression.

6. Experiments

This section describes some simple experiments on the well-known `Spambase` data set contributed by George Forman to the UCI Machine Learning Repository (Frank and Asuncion, 2010). Its overall size is 4601 examples and it contains examples of two classes: `email` (also written as 0) and `spam` (also written as 1). Hastie et al. (2009) report results of several machine-learning algorithms on this data set split randomly into a training set of size 3065 and test set of size 1536. The best result is achieved by MART (multiple additive regression tree; 4.5% error rate according to the second edition of Hastie et al. 2009).

We randomly permute the data set and divide it into 2602 examples for the proper training set, 999 for the calibration set, and 1000 for the test set. We consider the ICP whose conformity measure is defined by (1) where f is output by MART and

$$\Delta(y, f(x)) := \begin{cases} f(x) & \text{if } y = 1 \\ -f(x) & \text{if } y = 0. \end{cases} \tag{14}$$

MART's output $f(x)$ models the log-odds of `spam` vs `email`,

$$f(x) = \log \frac{P(1 \mid x)}{P(0 \mid x)},$$

which makes the interpretation of (14) as conformity score very natural.

The upper left plot in Figure 2 is the scatter plot of the pairs $(p^{\text{email}}, p^{\text{spam}})$ produced by the ICP for all examples in the test set. Email is shown as green noughts and spam as red crosses (and it is noticeable that the noughts were drawn after the crosses). The other two plots in the upper row are for email and spam separately. Ideally, email should be close to the horizontal axis and spam to the vertical axis; we can see that this is often true, with a few exceptions. The picture for the label conditional ICP looks almost identical: see the lower row of Figure 2.

Table 1 gives some statistics for the numbers of errors, multiple, and empty set predictions in the case of the (unconditional) ICP $\Gamma^{5\%}$ at significance level 5% (we obtain different numbers not only because of different splits but also because MART is randomized; the columns of the table correspond to the pseudorandom number generator seeds 0, 1, 2, etc.). The table demonstrates the validity, (lack of) conditional validity, and efficiency of the algorithm (the latter is of course inherited from the efficiency of MART). We give two kinds of conditional figures: the percentages of errors, multiple, and empty predictions for different labels and for two different kinds of objects. The two kinds of objects are obtained by splitting the object space $\mathbf{X}$ by the value of an attribute that we denote \$: it shows the percentage of the character \$ in the text of the message. The condition $\$ < 5.55\%$ was the root of the decision tree chosen both by Hastie et al. (2009, Section 9.2.5), who use all attributes in their analysis, and by Maindonald and Braun (2007, Chapter 11), who use 6 attributes chosen by them manually. (Both books use the `rpart` R package for decision trees.)

Notice that the numbers of errors, multiple predictions, and empty predictions tend to be greater for spam than for email. Somewhat counter-intuitively, they also tend to be greater for "email-like" objects containing few \$ characters than for "spam-like" objects. The percentage of multiple and empty predictions is relatively small since the error rate of the underlying predictor happens to be close to our significance level of 5%.

In practice, using a fixed significance level (such as the standard 5%) is not a good idea; we should at least pay attention to what happens at several significance levels. However, experimenting with prediction sets at a fixed significance level facilitates a comparison with theoretical results.

Table 2 gives similar statistics in the case of the label conditional ICP. The error rates are now about equal for email and spam, as expected. We refrain from giving similar predictable results for "object conditional" ICP with $\$ < 5.55\%$ and $\$ > 5.55\%$ as categories.

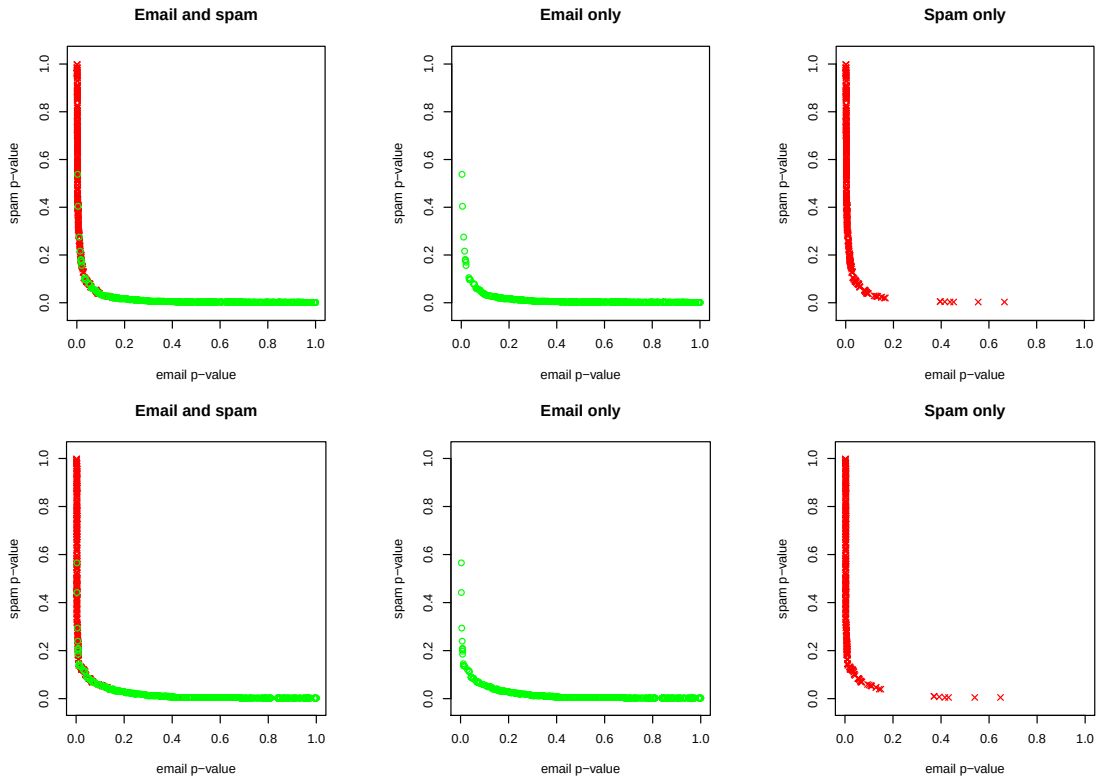

Figure 2: Scatter plots of the pairs $(p^{\text{email}}, p^{\text{spam}})$ for all examples in the test set (left plots), for email only (middle), and for spam only (right). The three upper plots are for the ICP and the three lower ones are for the label conditional ICP.

Figure 3 gives the calibration plots of the ICP for the test set. It shows approximate validity even for email and spam separately, except for the all-important lower-left corners. The latter are shown separately in Figure 4, where the lack of conditional validity becomes evident; cf. Figure 5 for the label conditional ICP.

From the numbers given in the "errors overall" row of Table 1 we can extract the corresponding confidence intervals for the probability of error conditional on the training set and MART's internal coin tosses; these are shown in Figure 6. It can be seen that training conditional validity is not grossly violated. (Notice that the 8 training sets used for producing this figure are not completely independent. Besides, the assumption of randomness might not be completely satisfied: permuting the data set ensures exchangeability but not necessarily randomness.) It is instructive to compare Figure 6 with the "theoretical" Figure 7 obtained from Propositions 2b (the thick blue line) and 2a (the thin red line). The dotted green line corresponds to the significance level 5%, and the black dot roughly corresponds to the maximal expected probability of error among 8 randomly chosen training sets. (It might appear that there is a discrepancy between Figures 6 and 7, but choosing different seeds usually leads to smaller numbers of errors than in Figure 6.)

RNG seed	0	1	2	3	4	5	6	7	Average
errors overall	4.1%	6.9%	4.6%	5.4%	5.3%	6.1%	7.7%	5.9%	5.75%
for email	2.44%	4.61%	2.26%	3.10%	4.49%	3.98%	5.02%	3.22%	3.64%
for spam	6.77%	10.43%	8.42%	9.02%	6.53%	9.32%	11.69%	10.29%	9.06%
for \$ $< 5.55\%$	4.36%	7.91%	5.15%	6.21%	6.27%	7.89%	8.79%	7.04%	6.70%
for \$ $> 5.55\%$	3.29%	4.12%	2.69%	2.64%	2.40%	1.13%	4.42%	2.15%	2.86%
multiple overall	2.7%	0%	0.1%	0%	0%	0.5%	0%	0%	0.41%
for email	2.11%	0%	0.16%	0%	0%	0.33%	0%	0%	0.33%
for spam	3.65%	0%	0%	0%	0%	0.76%	0%	0%	0.55%
for \$ $< 5.55\%$	3.04%	0%	0.13%	0%	0%	0.68%	0%	0%	0.48%
for \$ $> 5.55\%$	1.65%	0%	0%	0%	0%	0%	0%	0%	0.21%
empty overall	0%	2.7%	0%	1.2%	0.8%	0%	2.5%	0.4%	0.95%
for email	0%	1.48%	0%	0.65%	0.83%	0%	1.51%	0.64%	0.64%
for spam	0%	4.58%	0%	2.06%	0.75%	0%	3.98%	0%	1.42%
for \$ $< 5.55\%$	0%	3.14%	0%	1.55%	0.80%	0%	3.06%	0.52%	1.13%
for \$ $> 5.55\%$	0%	1.50%	0%	0%	0.80%	0%	0.80%	0%	0.39%

Table 1: Percentage of errors, multiple predictions, and empty predictions on the full test set and separately on email and spam. The results are given for various values of the seed for the R (pseudo)random number generator (RNG); column "Average" gives the average values for all 8 seeds 0–7.

RNG seed	0	1	2	3	4	5	6	7	Average
errors overall	3.4%	6.0%	3.8%	4.8%	5.7%	5.3%	6.5%	5.4%	5.11%
for email	3.73%	6.92%	3.87%	4.90%	6.64%	4.98%	5.85%	3.86%	5.10%
for spam	2.86%	4.58%	3.68%	4.64%	4.27%	5.79%	7.46%	7.92%	5.15%
multiple overall	4.2%	0%	4.0%	0%	0%	0.5%	0%	0.5%	1.15%
for email	3.90%	0%	5.48%	0%	0%	0.66%	0%	0.48%	1.32%
for spam	4.69%	0%	1.58%	0%	0%	0.25%	0%	0.53%	0.88%
empty overall	0%	1.0%	0%	0%	0.6%	0%	1.0%	0%	0.33%
for email	0%	1.48%	0%	0%	0.83%	0%	0.67%	0%	0.37%
for spam	0%	0.25%	0%	0%	0.25%	0%	1.49%	0%	0.25%

Table 2: The analogue of a subset of Table 1 in the case of the label conditional ICP.

7. ICPs and ROC curves

This section will discuss a close connection between an important class of ICPs ("probability-type" label conditional ICPs) and ROC curves. (For a previous study of connection between conformal prediction and ROC curves, see Vanderlooy and Sprinkhuizen-Kuyper 2007.) Let us say that an ICP or a label conditional ICP is *probability-type* if its inductive conformity measure is defined by (1) where f takes values in $\mathbb{R}$ and Δ is defined by (14).

The reader might have noticed that the two leftmost plots in Figure 2 look similar to a ROC curve. The following proposition will show that this is not coincidental in the case of the lower left one. However, before we state it, we need a few definitions. We will now consider a general binary classification problem and will denote the labels as 0 and 1. For

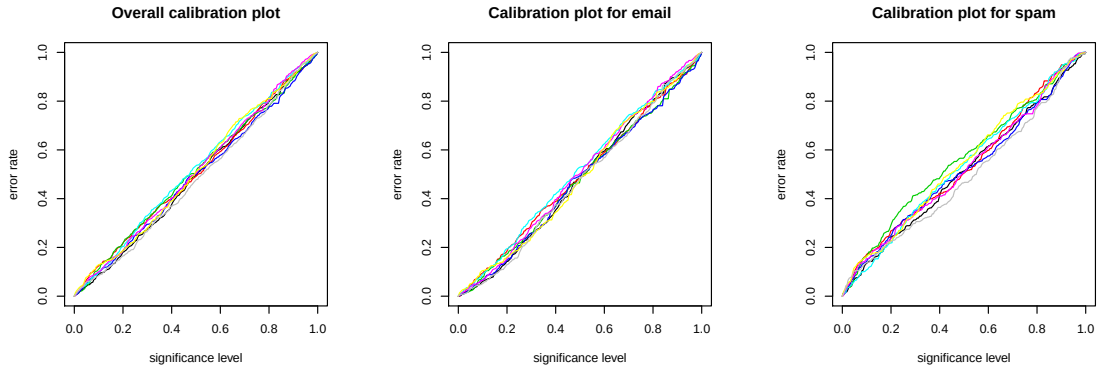

Figure 3: The calibration plot for the test set overall, the email in the test set, and the spam in the test set (for the first 8 seeds, 0–7).

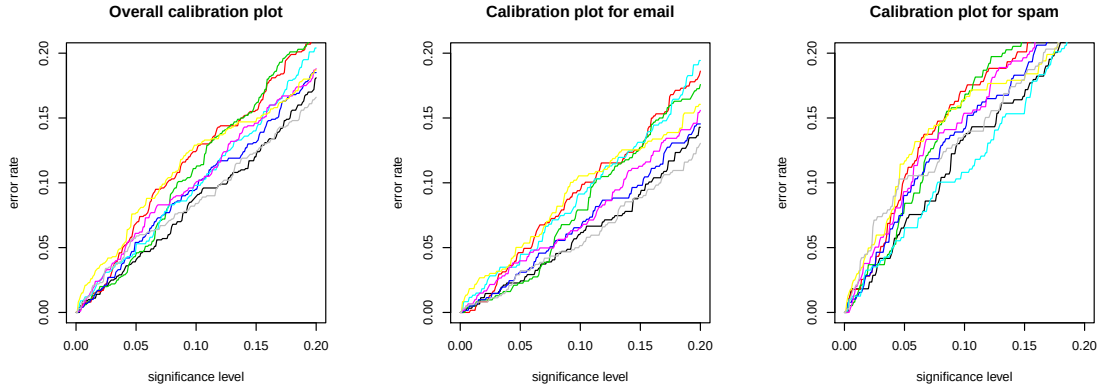

Figure 4: The lower left corners of the plots in Figure 3.

a threshold $c \in \mathbb{R}$, the *type I error on the calibration set* is

$$\alpha(c) := \frac{\{i = m+1, \dots, l \mid f(x_i) \geq c \,\&\, y_i = 0\}}{\{i = m+1, \dots, l \mid y_i = 0\}} \tag{15}$$

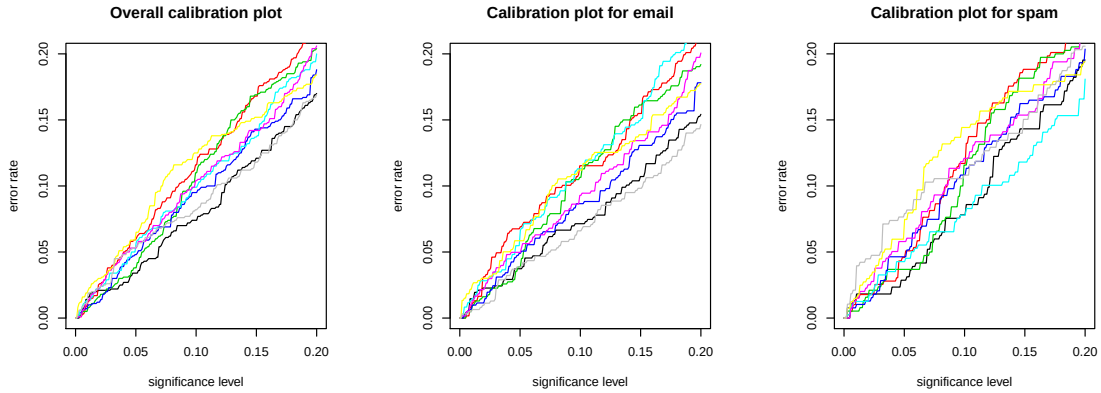

Figure 5: The analogue of Figure 4 for the label conditional ICP.

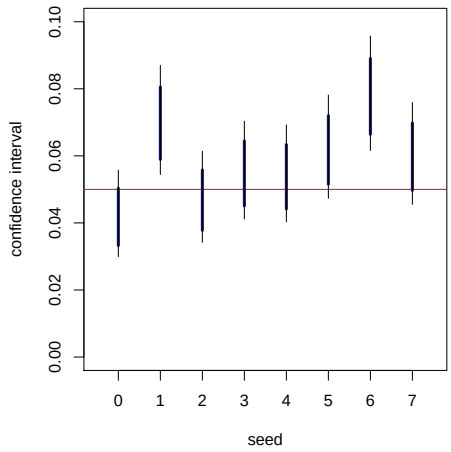

Figure 6: Confidence intervals for training conditional error probabilities: 95% in black (thin lines) and 80% in blue (thick lines). The 5% significance level is shown as the horizontal red line.

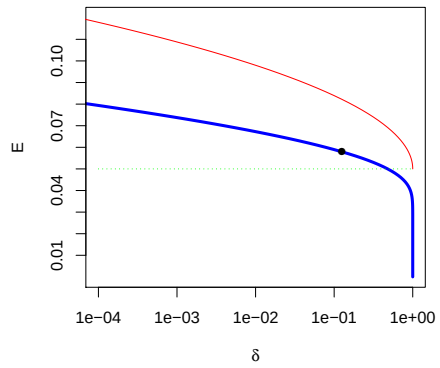

Figure 7: The probability of error E vs δ from Propositions 2b (the thick blue line) and 2a (the thin red line), where $\epsilon = 0.05$ and $n = 999$.

and the *type II error on the calibration set* is

$$\beta(c) := \frac{\{i = m+1, \dots, l \mid f(x_i) \le c \,\&\, y_i = 1\}}{\{i = m+1, \dots, l \mid y_i = 1\}} \tag{16}$$

(with 0/0 set, e.g., to 1/2). Intuitively, these are the error rates for the classifier that predicts 1 when $f(x) > c$ and predicts 0 when $f(x) < c$; our definition is conservative in that it counts the prediction as error whenever $f(x) = c$. The *ROC curve* is the parametric curve

$$\{(\alpha(c), \beta(c)) \mid c \in \mathbb{R}\} \subseteq [0,1]^2. \tag{17}$$

(Our version of ROC curves is the original version reflected in the line $y = 1/2$; in our sloppy terminology we follow Hastie et al. 2009, whose version is the original one reflected in the line $x = 1/2$, and many other books and papers.)

Proposition 5 *In the case of a probability-type label conditional ICP, for any object $x \in \mathbf{X}$, the distance between the pair (p^0, p^1) (see (8)) and the ROC curve is at most*

$$\sqrt{\frac{1}{(n^0+1)^2} + \frac{1}{(n^1+1)^2}}, \tag{18}$$

where n^y is the number of examples in the calibration set labelled as y.

Proof Let $c := f(x)$. Then we have

$$(p^0, p^1) = \left(\frac{n^0_{\geq} + 1}{n^0 + 1}, \frac{n^1_{\leq} + 1}{n^1 + 1} \right) \tag{19}$$

where $n^0_{\geq}$ is the number of examples (x_i, y_i) in the calibration set such that $y_i = 0$ and $f(x_i) \geq c$ and $n^1_{\leq}$ is the number of examples in the calibration set such that $y_i = 1$ and $f(x_i) \leq c$. It remains to notice that the point $\left(n^0_{\geq}/n^0, n^1_{\leq}/n^1\right)$ belongs to the ROC curve: the horizontal (resp. vertical) distance between this point and (19) does not exceed $1/(n^0+1)$ (resp. $1/(n^1+1)$), and the overall Euclidean distance does not exceed (18). ■

So far we have discussed the *empirical ROC curve*: (15) and (16) are the empirical probabilities of errors of the two types on the calibration set. It corresponds to the estimate k/n of the parameter of the binomial distribution based on observing k successes out of n. The minimax estimate is $(k + 1/2)/(n + 1)$, and the corresponding ROC curve (17) where $\alpha(c)$ and $\beta(c)$ are defined by (15) and (16) with the numerators increased by $\frac{1}{2}$ and the denominators increased by 1 will be called the *minimax ROC curve*. Notice that for the minimax ROC curve we can put a coefficient of $\frac{1}{2}$ in front of (18). Similarly, when using the Laplace estimate $(k + 1)/(n + 2)$, we obtain the *Laplace ROC curve*. See Figure 8 for the lower left corner of the lower left plot of Figure 2 with different ROC curves added to it.

In conclusion of our study of the `Spambase` data set, we will discuss the asymmetry of the two kinds of error in spam detection: classifying email as `spam` is much more harmful than letting occasional spam in. A reasonable approach is to start from a small number $\epsilon > 0$, the maximum tolerable percentage of email classified as `spam`, and then to try to minimize the percentage of spam classified as `email` under this constraint. The standard way of doing this is to classify a message x as `spam` if and only if $f(x) \geq c$, where c is the point on the ROC curve corresponding to the type I error ϵ. It is not clear what this means precisely, since we only have access to an estimate of the true ROC curve (and even on the true ROC curve such a point might not exist). But roughly, this means classifying x as `spam` if $f(x)$ exceeds the kth largest value in the set $\{\alpha_i \mid i \in \{m+1, \dots, l\} \,\&\, y_i = \texttt{email}\}$, where k is close to ϵn^0 and n^0 is the size of this set (i.e., the number of email in the calibration, or validation, set). To make this more precise, we can use the "one-sided label conditional ICP" classifying x as `spam` if and only if $p^0 \leq \epsilon$ for x. According to (19), this means that we classify x as `spam` if and only if $f(x)$ exceeds the kth largest value in the set $\{\alpha_i \mid i \in \{m + 1, \dots, l\} \,\&\, y_i = \texttt{email}\}$, where $k := \lfloor \epsilon(n^0 + 1) \rfloor$. The advantage of this version of the standard method is that it guarantees that the probability of mistaking email

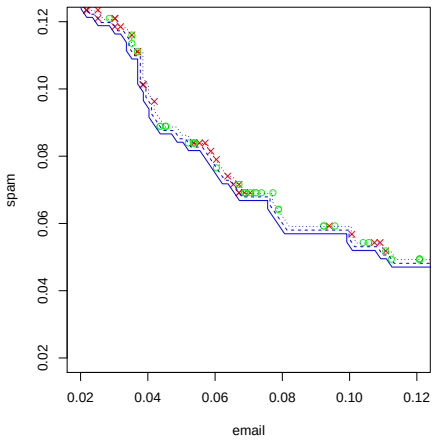

Figure 8: The lower left corner of the lower left plot of Figure 2 with the empirical (solid blue), minimax (dashed blue), and Laplace (dotted blue) ROC curves.

for spam is at most ϵ (see Proposition 3) and also enjoys the training conditional version of this property given by Proposition 2a (more accurately, its version for label conditional ICPs).

8. Conclusion

The goal of this paper has been to explore various versions of the requirement of conditional validity. With a small training set, we have to content ourselves with unconditional validity (or abandon any formal requirement of validity altogether). For bigger training sets training conditional validity will be approached by ICPs automatically, and we can approach example conditional validity by using conditional ICPs but making sure that the size of a typical category does not become too small (say, less than 100). In problems of binary classification, we can control false positive and false negative rates by using label conditional ICPs.

The known property of validity of inductive conformal predictors (Proposition 1) can be stated in the traditional statistical language (see, e.g., Fraser 1957) by saying that they are $1-\epsilon$ expectation tolerance regions, where ϵ is the significance level. In classical statistics, however, there are two kinds of tolerance regions: $1-\epsilon$ expectation tolerance regions and PAC-type $1-\delta$ tolerance regions for a proportion $1-\epsilon$, in the terminology of Fraser (1957). We have seen (Proposition 2a) that inductive conformal predictors are tolerance regions in the second sense as well (cf. Vovk 2012, Appendix A).

A disadvantage of inductive conformal predictors is their potential predictive inefficiency: indeed, the calibration set is wasted as far as the development of the prediction rule f in (1) is concerned, and the proper training set is wasted as far as the calibration (3) of conformity scores into p-values is concerned. Conformal predictors use the full training set for both purposes, and so can be expected to be significantly more efficient. (There have been reports of comparable and even better predictive efficiency of ICPs as compared to conformal predictors but they may be unusual artefacts of the methods used and particular data sets.) It is an open question whether we can guarantee training conditional validity

under (5) or a similar condition for conformal predictors different from classical tolerance regions. Perhaps no universal results of this kind exist, and different families of conformal predictors will require different methods.

Acknowledgments

The empirical studies described in this paper used the R system and the `gbm` package written by Greg Ridgeway. This work was partially supported by the Cyprus Research Promotion Foundation. Many thanks to the reviewers for their advice.

References

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Donald A. S. Fraser. *Nonparametric Methods in Statistics*. Wiley, New York, 1957.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 2009.

Jing Lei and Larry Wasserman. Distribution free prediction bands. Technical Report `arXiv:1203.5422 [stat.ME]`, `arXiv.org` e-Print archive, March 2012.

Jing Lei, James Robins, and Larry Wasserman. Efficient nonparametric conformal prediction regions. Technical Report `arXiv:1111.1418 [math.ST]`, `arXiv.org` e-Print archive, November 2011.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. Generalized conformal prediction for functional data. 2012.

Jon Maindonald and John Braun. *Data Analysis and Graphics Using R: An Example-Based Approach*. Cambridge University Press, Cambridge, second edition, 2007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2010.

Stijn Vanderlooy and Ida G. Sprinkhuizen-Kuyper. A comparison of two approaches to classify with guaranteed performance. In *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 4702 of *Lecture Notes in Computer Science*, pages 288–299, Berlin, 2007. Springer.

Vladimir Vovk. Conditional validity of inductive conformal predictors. Technical Report `arXiv:1209.2673` [cs.LG], `arXiv.org` e-Print archive, September 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.