Cross-conformal predictors

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #6

First posted August 5, 2012. Last revised January 10, 2013.

Project web site:
http://alrw.net

Abstract

Inductive conformal predictors have been designed to overcome the computational inefficiency exhibited by conformal predictors for many underlying prediction algorithms. Whereas computationally efficient, inductive conformal predictors sacrifice different parts of the training set at different stages of prediction, which affects their informational efficiency. This paper introduces the method of cross-conformal prediction, which is a hybrid of the methods of inductive conformal prediction and cross-validation, and studies its validity and informational efficiency empirically. The computational efficiency of cross-conformal predictors is comparable to that of inductive conformal predictors, and they produce valid predictions in our empirical studies.

Contents

1 Introduction

The method of conformal prediction produces set predictions that are automatically valid in the sense of their unconditional coverage probability being equal to or exceeding a preset confidence level ([19], Chapter 2). A more computationally efficient method of this kind is that of inductive conformal prediction ([14]; [19], Section 4.1; [1]). However, inductive conformal predictors are typically less informationally efficient, in the sense of producing larger prediction sets as compared with conformal predictors. Motivated by the method of cross-validation [13, 16], this paper explores a hybrid method, which we call cross-conformal prediction.

We are mainly interested in the problems of classification and regression, in which we are given a training set consisting of examples, each example consisting of an object and a label, and asked to predict the label of a new test object; in the problem of classification labels are elements of a given finite set, and in the problem of regression labels are real numbers. (Our experimental results will involve only classification problems.) If we are asked to predict labels for more than one test object, the same prediction procedure can be applied to each test object separately. In this introductory section and in most of our empirical studies we consider the problem of binary classification, in which labels can take only two values, which we will encode as 0 and 1. We always assume that the examples (both the training examples and the test examples, consisting of given objects and unknown labels) are generated from an exchangeable probability measure (i.e., a probability measure that is invariant under permuting the examples). This *exchangeability assumption* is slightly weaker than the *assumption of randomness* that the examples are generated independently from the same probability measure.

The idea of conformal prediction is to try the two different labels, 0 and 1, for the test object, and for either postulated label to test the assumption of exchangeability by checking how well the test example conforms to the training set; the output of the procedure is the corresponding p-values p^0 and p^1. Two standard ways to package the pair (p_0, p_1) are:

- Report the *predicted label* $\arg\max_{y\in\{0,1\}} p^y$, *confidence* $1 - \min(p^0, p^1)$, and *credibility* $\max(p^0, p^1)$.
- For a given significance level $\epsilon \in (0, 1)$ output the corresponding prediction set $\{y \mid p^y > \epsilon\}$.

The prediction sets output by conformal predictors make an error, i.e., fail to cover the true label, with probability at most ϵ. In empirical studies this shows as the calibration plot (the plot of the percentage of errors against $\epsilon \in (0, 1)$) being below the bisector of the first quadrant, to within statistical fluctuations; in practice, calibration plots are usually very close to the bisector of the first quadrant.

In inductive conformal prediction, discussed in Section 2 of this paper, the training set is split into two parts, the proper training set and the calibration

set. The two p-values p^0 and p^1 are computed by checking how well the test example conforms to the calibration set. The way of checking conformity is based on a prediction rule found from the proper training set and produces, for each example in the calibration set and for the test example, the corresponding "conformity score". The conformity score of the test example is then calibrated to the conformity scores of the calibration set to obtain the p-value. For details, see Section 2.

Inductive conformal predictors are usually much more computationally efficient than the corresponding conformal predictors (also discussed in Section 2). However, they are less informationally efficient: they use only the proper training set when developing the prediction rule and only the calibration set when calibrating the conformity score of the test example, whereas conformal predictors use the full training set for both purposes.

Cross-conformal prediction (Section 3) modifies inductive conformal prediction in order to use the full training set for calibration and significant parts of the training set (such as 80% or 90%) for developing prediction rules. The training set is split into K folds of equal (or almost equal) size; in our experiments we use $K = 5$ or $K = 10$. For each $k = 1, \dots, K$ we construct a separate inductive conformal predictor using the kth fold as the calibration set and the rest of the training set as the proper training set. Let (p_k^0, p_k^1) be the corresponding p-values. Next the two sets of p-values, p_k^0 and p_k^1, are merged into combined p-values p^0 and p^1, which are the result of the procedure. In the method of cross-conformal prediction we, essentially, combine p-values by averaging them.

Empirical studies reported in Section 3 show that cross-conformal predictors are valid in the sense of their calibration plots being close to the bisector of the first quadrant (in this case we also say that their predictions are well calibrated). In our empirical studies in this paper we mainly use the well-known Spambase data set. The underlying algorithm that we use is Freedman's gradient boosting (also known as MART), which performs particularly well on the Spambase data set [10]; however, because of its computational inefficiency, it is utterly infeasible to use it in combination with conformal prediction. We use the same data set to demonstrate the efficiency of cross-conformal predictors as compared with inductive conformal predictors.

Besides the Spambase data set we use another well-known dataset, the USPS data set of hand-written digits, in combination with the 1-Nearest Neighbour algorithm for tangent distance, which is one of the best performing algorithms on this data set. Now it becomes computationally feasible to use conformal prediction, and we show that cross-conformal prediction works almost as well as conformal prediction. Our experiments on the USPS data set also confirm the empirical validity of cross-conformal predictors and their greater efficiency as compared with inductive conformal predictors.

Inductive conformal predictors guarantee, and Section 3 studies empirically, the notion of validity that we call unconditional validity since it is a guarantee on unconditional coverage probability. Section 4 introduces a conditional version of cross-conformal predictors and studies empirically its conditional validity. Section 5 concludes.

Appendix A discusses the intuition behind an extreme case of cross-conformal prediction that we call leave-one-out conformal prediction. It explains why cross-conformal predictors do not enjoy the same theoretical guarantee of unconditional validity as inductive conformal predictors.

Appendix B reports results of empirical studies on the Spambase data set of two bootstrap versions of conformal prediction. The basic version is empirically valid but somewhat less efficient than cross-conformal predictors. The randomized version is not empirically valid.

In Appendix C we consider the method of cross-conformal prediction in which averaging p-values is replaced by the most standard way of combining p-values, Fisher's method [5]. However, this method produces badly miscalibrated results. This is not surprising, since Fisher's method assumes the independence of the p-values being combined, whereas in this case they are heavily dependent.

2 Conformal predictors and inductive conformal predictors

We fix two measurable spaces: $\mathbf{X}$, called the *object space*, and $\mathbf{Y}$, called the *label space*. The Cartesian product $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is the *example space*. A *training set* is a sequence $(z_1, \dots, z_l) \in \mathbf{Z}^l$ of *examples* $z_i = (x_i, y_i)$, where $x_i \in \mathbf{X}$ are the *objects* and $y_i \in \mathbf{Y}$ are the *labels*.

A *conformity measure* is a measurable function $A : \mathbf{Z}^* \times \mathbf{Z} \to \mathbb{R}$ such that $A(\zeta, z)$ does not depend on the order of the elements of $\zeta \in \mathbf{Z}^*$. The idea behind the *conformity score* $A(\zeta, z)$ is that it should measure how well the example z conforms to the examples in the sequence ζ. The *conformal predictor* (CP) corresponding to A is defined as the set predictor

$$\Gamma^\epsilon(z_1, \dots, z_l, x) := \{y \mid p^y > \epsilon\}, \tag{1}$$

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

$$p^y := \frac{|\{i \in \{1, \dots, l\} \mid \alpha_i \le \alpha^y\}| + 1}{l + 1}, \tag{2}$$

and

$$\alpha_i := A((z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_l, z), z_i), \quad i \in \{1, \dots, l\},$$
$$\alpha^y := A((z_1, \dots, z_l), (x, y)) \tag{3}$$

are the conformity scores of the training and test examples. Given the training set and a test object x the CP predicts its label y; it *makes an error* (at significance level ϵ) if $y \notin \Gamma^\epsilon(z_1, \dots, z_l, x)$.

In this paper we will use the *1-Nearest Neighbour conformity measure*

$$A(((x_1, y_1), \dots, (x_n, y_n)), (x, y)) := \frac{\min_{i=1,\dots,n:y_i \ne y} d(x, x_i)}{\min_{i=1,\dots,n:y_i = y} d(x, x_i)}, \tag{4}$$

where $d : \mathbf{X}^2 \to [0, \infty)$ is a measure of distance between two points (often, but not necessarily, a metric). The intuition behind (4) is that an example conforms to a set of examples if it is much closer to an example in the set with the same label than to any example with a different label.

For $S \subseteq \{1, \dots, l\}$, we let z_S stand for the sequence $(z_{s_1}, \dots, z_{s_n})$, where $s_1, \dots, s_n$ is the sequence of all elements of S listed in the increasing order (so that $n := |S|$).

In the method of inductive conformal prediction, we split the training set into two non-empty parts, the *proper training set* z_T and the *calibration set* z_C, where (T, C) is a partition of $\{1, \dots, l\}$. An *inductive conformity measure* is a measurable function $A : \mathbf{Z}^* \times \mathbf{Z} \to \mathbb{R}$. We are only interested in the case where $A(\zeta, z)$ does not depend on the order of the elements of $\zeta \in \mathbf{Z}^*$, albeit this is not part of the definition. (In particular, in our empirical studies we will never use inductive conformity measures that are not conformity measures.) The *conformity score* $A(z_T, z)$ will be used to measure how well the example z conforms to the proper training set z_T. A standard choice is

$$A(z_T, (x, y)) := \Delta(y, f(x)), \tag{5}$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is a prediction rule found from z_T 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 $\mathbf{Y}$ (usually $\mathbf{Y}' \supset \mathbf{Y}$) may be useful when the underlying prediction method gives additional information to the predicted label; e.g., the MART procedure used in this paper gives the logit of the predicted probability that the label is 1.

The *inductive conformal predictor* (ICP) corresponding to an inductive conformity measure A is defined as the set predictor (1), where the p-values p^y, $y \in \mathbf{Y}$, are now defined by

$$p^y := \frac{|\{i \in C \mid \alpha_i \le \alpha^y\}| + 1}{|C| + 1}, \tag{6}$$

and the conformity scores are

$$\alpha_i := A(z_T, z_i), \quad i \in C, \qquad \alpha^y := A(z_T, (x, y)). \tag{7}$$

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

Proposition 1 ([19], Propositions 2.3 and 4.1)**.** *Let Γ be a conformal predictor or an inductive conformal predictor. If random examples $Z_1, \dots, Z_l, Z = (X, Y)$ are exchangeable (i.e., their distribution is invariant under their permutations), the probability of error $Y \notin \Gamma^\epsilon(Z_1, \dots, Z_l, X)$ does not exceed ϵ for any ϵ.*

We call the property of conformal predictors and inductive conformal predictors asserted in Proposition 1 unconditional validity since it is about the unconditional probability of error; we often abbreviate it to "validity" omitting

"unconditional". Various conditional properties of validity are discussed in [12] and, in more detail, [20].

To check the validity of a family (Γ^ϵ) of set predictors, such as a conformal predictor or an inductive conformal predictor, empirically on given training and test sets one can use the *calibration plot*: the function mapping each significance level ϵ to the percentage of erroneous predictions made by the set predictor Γ^ϵ on the test set. However, Proposition 1 does not guarantee that at each significance level ϵ with high probability the calibration plot of a conformal predictor or an inductive conformal predictor will be close to or below the bisector of the first quadrant: errors on different test examples are not independent, since predictions are computed from the same training set.

Figure 1 shows the calibration plots for a conformal predictor and an inductive conformal predictor on the USPS data set. The data set, which consists of 9298 labelled hand-written images, has been divided randomly into a training set of size 7200 and a test set of size 2098 (we cannot use the original split into the training and test sets as it violates the exchangeability assumption: see, e.g., [19], Section 7.1). The conformity measure used for both the CP and the ICP is (4) with d tangent distance [15]. In the case of the ICP, the training set is randomly divided into a proper training set and a calibration set in proportion 2:1, as discussed below. The experiments are repeated 8 times to get an idea of how much their results are affected by the random splits.

The plots in Figure 1 indicate that both kinds of predictors are empirically well calibrated (this phenomenon was first observed in [17]). In the case of inductive conformal predictors a theoretical explanation can be found in [20] (Proposition 2a): with a high probability, the conditional probability of error given the training set will be close to ϵ. Since errors on different test examples are conditionally independent given the training set, this implies good calibration: at each significance level ϵ with high probability the calibration plot of an inductive conformal predictor will be close to or below the bisector of the first quadrant.

The family of prediction sets $\Gamma^\epsilon(z_1, \dots, z_l, x)$, $\epsilon \in (0, 1)$, is just one possible way of packaging the p-values p^y. Another way, already discussed in Section 1 in the context of binary classification, is to report the *predicted label* $\arg\max_{y \in \mathbf{Y}} p^y$, *confidence* $1 - p$, where p is the second largest p-value among p^y, and the *credibility* $\max_y p^y$. In the case of binary classification the predicted confidence and credibility carry the same information as the full set $\{p^y \mid y \in \mathbf{Y}\}$ of p-values, but this is not true in general. It is clear that the notion of confidence is likely to be useful only in classification problems.

In our experiments reported in this and the following sections we split the training set into the proper training set and the calibration set in proportion $2 : 1$. This is the most standard proportion (cf. [10], p. 222, where the validation set plays a similar role to our calibration set), but the ideal proportion depends on the learning curve for the given problem of prediction (cf. [10], Figure 7.8). Too small a calibration set leads to a high variance of confidence (since calibrating conformity scores becomes unreliable) and too small a proper training set leads to a downward bias in confidence (conformity scores based on a small proper

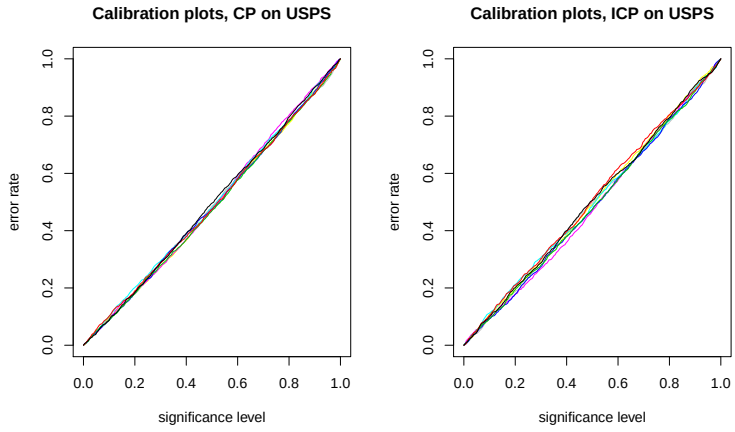

Figure 1: Left panel: the calibration plots on the USPS data set for the conformal predictor and the first 8 seeds, 0–7, for the pseudorandom number generator. Right panel: the analogous plots for the inductive conformal predictor.

training set cannot produce confident predictions). In the next section we will see that using cross-conformal predictors improves both bias and variance (cf. Table 1).

3 Cross-conformal predictors

Cross-conformal predictors (CCPs) are defined as follows. The training set is split into K non-empty subsets (*folds*) z_{S_k}, $k = 1, \dots, K$, where $K \in \{2, 3, \dots\}$ is a parameter of the algorithm and $(S_1, \dots, S_K)$ is a partition of $\{1, \dots, l\}$. For each $k \in \{1, \dots, K\}$ and each potential label $y \in \mathbf{Y}$ of the test object x find the conformity scores of the examples in z_{S_k} and of (x, y) by

$$\alpha_{i,k} := A(z_{S_{-k}}, z_i), \quad i \in S_k, \qquad \alpha_k^y := A(z_{S_{-k}}, (x, y)), \tag{8}$$

where $S_{-k} := \cup_{j \neq k} S_j$ and A is a given inductive conformity measure. The corresponding p-values are defined by

$$p^y := \frac{\sum_{k=1}^K |\{i \in S_k \mid \alpha_{i,k} \leq \alpha_k^y\}| + 1}{l + 1}. \tag{9}$$

Confidence and credibility are now defined as before; the set predictor Γ^ϵ is also defined as before, by (1), where the significance level $\epsilon > 0$ is another parameter.

The definition of CCPs parallels that of ICPs, except that we now use the whole training set for calibration. The conformity scores (8) are computed as in (7) but using the current fold as the calibration set and the union of all the folds except for the current one as the proper training set. Calibration (9) is

done by combining the ranks of the test example (x, y) with a postulated label in all the folds.

If we define the separate p-value

$$p_k^y := \frac{|\{i \in S_k \mid \alpha_{i,k} \le \alpha_k^y\}| + 1}{|S_k| + 1} \tag{10}$$

for each fold, we can see that p^y is essentially the average of p_k^y. In particular, if each fold has the same size, $|S_1| = \cdots = |S_K|$, a simple calculation gives

$$p^y = \bar{p}^y + \frac{K-1}{l+1}(\bar{p}^y - 1) \approx \bar{p}^y, \tag{11}$$

where $\bar{p}^y := \frac{1}{K}\sum_{k=1}^K p_k^y$ is the arithmetic mean of p_k^y and the $\approx$ assumes $K \ll l$.

In this paper we give calibration plots for 10-fold cross-conformal prediction; calibration plots for 5-fold cross-conformal prediction are part of Online Resource 1. We take $K \in \{5, 10\}$ following the advice for cross-validation in [10], who refer to Breiman and Spector [2] and Kohavi [11]. (Our setting, however, is somewhat different from cross-validation, and it is not obvious whether $K \in \{5, 10\}$ remains a good choice.) In the experiments of this section we use the Spambase and USPS data sets. The size of the Spambase data set is 4601, and there are two labels: `email`, encoded as 0, and `spam`, encoded as 1. The USPS data set is bigger, 9298 examples, and multilabel (0–9).

The conformity measure used in the case of the Spambase data set is (5), where f is output by MART ([10], Chapter 10) and

$$\Delta(y, f(x)) := \begin{cases} f(x) & \text{if } y = 1 \\ -f(x) & \text{if } y = 0. \end{cases} \tag{12}$$

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 (12) as conformity score very natural. In the case of the USPS data set, we always use the conformity measure (4) with d tangent distance.

The R and MATLAB programs used in the experiments described in this paper have been uploaded to `arXiv` (see [18]). The R programs, used for processing the Spambase data set, rely on the `gbm` package with virtually all parameters set to the default values (given in the description provided in response to `help("gbm")`). The only parameter that has been modified is `n.trees`, the number of trees, which should be as large as possible and whose default value was clearly insufficient. The MATLAB programs, used for processing the USPS data set, rely on the C program for computing tangent distance (with one-sided distance and all tangents) written by Daniel Keysers.

Figure 2 (the two top panels) gives the calibration plots for the CCP and for 8 random splits of the data sets into a training set (of size 3600 for Spambase and

Table 1: Mean (over the test set) confidence and credibility for the ICP and the 5-fold and 10-fold CCP on the Spambase data set. The results are given for various values of the seed for the pseudorandom number generator; column "Average" gives the average of all the 100 values for the seeds 0–99, and column "St. dev." gives the estimate of the standard deviation computed from those 100 values.

Seed	0	1	...	99	Average	St. dev.
mean confidence, ICP	99.25%	99.23%	...	99.14%	99.158%	0.149%
mean credibility, ICP	51.31%	50.38%	...	51.44%	50.922%	1.144%
mean confidence, $K = 5$	99.22%	99.17%	...	99.28%	99.232%	0.061%
mean credibility, $K = 5$	51.11%	49.75%	...	50.78%	50.745%	0.910%
mean confidence, $K = 10$	99.24%	99.20%	...	99.31%	99.253%	0.055%
mean credibility, $K = 10$	51.08%	49.74%	...	50.70%	50.735%	0.928%

7200 for USPS) and a test set (of size 1001 for Spambase and 2098 for USPS) and of the training set into 10 folds of equal size. In the case of Spambase, there is a further source of randomness as the MART procedure is itself randomized. The two bottom panels of Figure 2 give the lower left corners of the plots in the top panels: these are the most important parts of calibration plots in applications. The analogous plots for 5 folds are given in Online Resource 1. Visually, all plots are well calibrated (close to the bisector of the first quadrant). Since the USPS data set is bigger, the corresponding plots are closer to the bisector of the first quadrant.

As for the efficiency of the CCP on the Spambase data set, see Table 1, which gives some statistics for the confidence and credibility output by the ICP (with the 2 : 1 split into the proper training and calibration sets, as already mentioned) and the 5-fold and 10-fold CCP. The columns labelled "0" to "99" give the mean values of confidence and credibility over the test set for various values of the seed for the pseudorandom number generator. The column labelled "Average" gives the average $\bar{v} := \frac{1}{100}\sum_{i=0}^{99} v_i$ of all the 100 values (which we denote $v_0, \dots, v_{99}$) for the seeds 0 to 99, and the column labelled "St. dev." gives the estimate $(\frac{1}{99}\sum_{i=0}^{99}(v_i - \bar{v})^2)^{1/2}$ of the standard deviation of the mean values computed from $v_0, \dots, v_{99}$ (the square root of the standard unbiased estimate of the variance of the mean values). The advantage of the CCP over the ICP that can be seen from the table is that it gives higher and more stable mean confidence values: see the last two columns.

Similar results for the USPS data set are given in Table 2. This table, however, also contains information about the CP, which becomes feasible in the case of the 1-Nearest Neighbour underlying algorithm that we use for the USPS data set. The CCP is almost as efficient as the CP (especially in the case of 10 folds) and significantly more efficient than the ICP.

Table 2: Mean confidence and credibility for the ICP, 5- and 10-fold CCP, and CP on the USPS data set. The results are given for various values of the seed for the pseudorandom number generator; columns "Average" and "St. dev." give the averages and estimates of standard deviations as in Table 1.

Seed	0	1	...	99	Average	St. dev.
mean confidence, ICP	99.85%	99.79%	...	99.79%	99.823%	0.044%
mean credibility, ICP	50.31%	49.72%	...	51.15%	50.135%	0.932%
mean confidence, $K = 5$	99.88%	99.85%	...	99.85%	99.846%	0.018%
mean credibility, $K = 5$	50.39%	50.40%	...	50.79%	50.059%	0.748%
mean confidence, $K = 10$	99.90%	99.87%	...	99.86%	99.855%	0.017%
mean credibility, $K = 10$	50.39%	50.30%	...	50.82%	50.045%	0.757%
mean confidence, CP	99.91%	99.87%	...	99.87%	99.860%	0.017%
mean credibility, CP	50.92%	51.46%	...	51.44%	50.893%	0.755%

4 Conditional cross-conformal predictors

There are several natural kinds of conditional validity for set predictors: see, e.g., [20], Figure 1. Achieving some of these kinds (such as label and object conditional validity, in the terminology of [20]) requires modifying the definition of conformal predictors. Another kind (training conditional validity) is achieved automatically, at least in some cases: see [20], Section 3. In this paper we only discuss the label conditional version of conformal predictors and their variants, which ensures the validity conditional on the label of the test example. As will be discussed later, the property of label conditional validity is particularly important in applications where different kinds of errors have different significance, such as automatic spam filtering.

The only difference of *label conditional conformal predictors* from CPs is that (2) is replaced by

$$p^y := \frac{|\{i \in \{1,\dots,l\} \mid y_i = y \,\&\, \alpha_i \le \alpha^y\}| + 1}{|\{i \in \{1,\dots,l\} \mid y_i = y\}| + 1}.$$

And the only difference of *label conditional inductive conformal predictors* from ICPs is that (6) is replaced by

$$p^y := \frac{|\{i \in C \mid y_i = y \,\&\, \alpha_i \le \alpha^y\}| + 1}{|\{i \in C \mid y_i = y\}| + 1}.$$

The following proposition is the label conditional version of Proposition 1.

Proposition 2 ([19], Proposition 4.10)**.** *Let Γ be a label conditional conformal predictor or a label conditional inductive conformal predictor. If random examples $Z_1,\dots,Z_l$, $Z = (X,Y)$ are exchangeable, the conditional probability of error $Y \notin \Gamma^\epsilon(Z_1,\dots,Z_l,X)$ given Y does not exceed ϵ for any ϵ.*

Label conditional cross-conformal predictors (abbreviated to CCCP, as this is the only kind of conditional CCP that we consider) are defined as CCPs except that (9) is replaced by

$$p^y := \frac{\sum_{k=1}^K |\{i \in S_k \mid y_i = y \,\&\, \alpha_{i,k} \le \alpha_k^y\}| + 1}{\sum_{k=1}^K |\{i \in S_k \mid y_i = y\}| + 1}.$$

First we check empirically whether CCCPs are well calibrated. Figure 3 shows separate calibration plots for the test examples labelled as `email` and `spam`, both for $K = 10$ folds. Both plots are close to the bisector of the first quadrant. This is important as we are not really interested in the overall error rate: for example, when using the predictor for spam filtering, first of all we want to get the amount of email classified as `spam` down to an admissible low level, and only after that we try to optimize the amount of spam classified as `email`. Moreover, in the case of label conditional predictors nothing prevents us from having different significance levels for `email` and `spam`: we can replace the ϵ in (1) by ϵ^y allowing the confidence level ϵ^y to depend on the label $y \in \mathbf{Y}$.

Figure 4 gives the scatter plot of the pairs (p^0, p^1) for all test examples, where p^0 is the p-value when the example is labelled as `email` and p^1 is the p-value when it is labelled as `spam`. The following two tables (Table 3 and Table 4) will give some summary information for the data represented in this figure. It has been shown in [20] (see, e.g., Figure 8) that in the case of ICPs there is a close connection between scatter plots of p-values and empirical ROC curves. Figure 4 (and especially its right panel) shows that there are no similar close connections in the case of CCPs.

Table 3 shows the "confusion matrices" for `email` and `spam` for the first 100 seeds of the pseudorandom number generator. It shows the mean p-values for email in the test set when classified as `email`, for email in the test set when classified as `spam`, for spam in the test set when classified as `email`, and for spam in the test set when classified as `spam`. The p-values for email when classified as `email` and for spam when classified as `spam` are distributed approximately uniformly in the interval $[0, 1]$, and so their means should be approximately 50%; this is what Table 3 shows, confirming the approximate validity of the method. The p-values for email when classified as `spam` and for spam when classified as `email` should be small for efficient prediction methods, and we can see that indeed they never exceed 2% in Table 3 (and very rarely exceed 2% if the table is expanded by adding the missing values for the seeds 2–98).

Table 4 demonstrates the validity and efficiency of spam filters based on 10-fold CCCPs with target probabilities 1%, 2%, and 5% of mistaking email for spam. For a given target probability $\epsilon \in \{0.01, 0.02, 0.05\}$ the spam filter classifies a test object as `spam` if and only if $p^0 \le \epsilon$ (ignoring p^1; remember that 0 encodes `email` and 1 `spam`). The table confirms the validity of the CCCP conditional on the label being `email` and gives an indication of its efficiency (the amount of spam let in).

Figure 5 gives similar information about validity and efficiency for all target probabilities $\epsilon \in (0, 1)$: the left panels confirm the label conditional validity

Table 3: Mean (over the test set) p-values for email if classified as `email`, for email if classified as `spam`, for spam if classified as `email`, and for spam if classified as `spam`. The results are given for 100 values of the seed for the pseudorandom number generator; column "Average" gives the averages of the means over the 100 seeds 0–99, and column "St. dev." gives the estimates of the standard deviations of the means.

Seed	0	1	...	99	Average	St. dev.
email as `email`	50.24%	48.61%	...	49.41%	50.021%	1.133%
email as `spam`	1.44%	1.38%	...	1.45%	1.590%	0.206%
spam as `email`	1.66%	1.56%	...	1.63%	1.591%	0.251%
spam as `spam`	50.68%	49.93%	...	50.97%	50.044%	1.644%

Table 4: The percentage of misclassified email and spam in the test set for the spam filters based on conditional cross-conformal prediction. The results are given for various values of the seed for the pseudorandom number generator; the last two columns give the averages and estimates of standard deviations.

Seed	0	1	...	99	Average	St. dev.
email at 1%	0.97%	0%	...	1.16%	0.999%	0.443%
spam at 1%	23.64%	26.14%	...	16.88%	20.275%	2.582%
email at 2%	1.79%	1.81%	...	1.82%	1.993%	0.617%
spam at 2%	12.47%	12.94%	...	10.08%	12.211%	1.620%
email at 5%	4.55%	5.11%	...	4.30%	4.861%	0.889%
spam at 5%	4.68%	5.08%	...	5.04%	5.543%	0.954%

of our spam filters and the right panels can be regarded as measuring their efficiency. (Notice that the left panels of Figures 3 and 5 are identical.)

5 Conclusion

Conformal prediction and inductive conformal prediction are two approaches to the theory of tolerance regions (see, e.g., [6]). The known validity results for conformal and inductive conformal predictors can be expressed by saying that they are $1 - \epsilon$ expectation tolerance regions, where ϵ is the significance level (see Proposition 1 above). It is also known ([20], Proposition 2a) that inductive conformal predictors are $1-\delta$ tolerance regions for a proportion $1-\epsilon$ for suitable δ and ϵ. On the other hand, at this time there are no theoretical results about the validity of cross-conformal predictors, and it is an interesting open problem to establish such results.

Acknowledgments

The empirical studies described in this paper used the R system, MATLAB, the R package `gbm` written by Greg Ridgeway (based on the work of Freund and Schapire [7] and Friedman [8, 9]), and the C program for computing tangent distance written by Daniel Keysers and adapted to MATLAB by Aditi Krishn. An extended abstract of an early version of this paper was published in the Proceedings of the Fifth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2012, Amsterdam, August 2012, `http://event.cwi.nl/witmse2012/proc.pdf`). I am grateful to participants in WITMSE 2012 and COPA 2012 (First Workshop on Conformal Prediction and its Applications, Halkidiki, Greece, September 2012) for useful discussions. In particular, Martin Eklund's comments were helpful in improving presentation. This work was partially supported by the Cyprus Research Promotion Foundation.

References

[1] Anonymous. Generalized conformal prediction for functional data. Submitted for publication, June 2012.

[2] Leo Breiman and Philip Spector. Submodel selection and evaluation in regression: the X-random case. *International Statistical Review*, 60:291–319, 1992.

[3] Bradley Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[4] Bradley Efron. Estimating the error rate of a prediction rule: some improvements on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[5] Ronald A. Fisher. Combining independent tests of significance. *American Statistician*, 2:30, 1948. This is the answer to Question 14 in Frederick Mosteller's "Questions and Answers" column.

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

[7] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[8] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[9] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2002.

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

[11] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1143, San Mateo, CA, 1995. Morgan Kaufmann.

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

[13] Frederick Mosteller and John W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, volume 2, pages 80–203. Addison-Wesley, Reading, MA, second edition, 1968.

[14] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the First International Conference on Machine Learning and Applications (ICMLA)*, pages 159–163, Las Vegas, NV, 2002. CSREA Press.

[15] Patrice Simard, Yann LeCun, and John Denker. Efficient pattern recognition using a new transformation distance. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 50–58, San Mateo, CA, 1993. Morgan Kaufmann.

[16] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

[17] Stijn Vanderlooy, Laurens van der Maaten, and Ida Sprinkhuizen-Kuyper. Off-line learning with Transductive Confidence Machines: an empirical evaluation. In Petra Perner, editor, *Proceedings of the Fifth International Conference on Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Artificial Intelligence*, pages 310–323, Berlin, 2007. Springer.

[18] Vladimir Vovk. Cross-conformal predictors. Technical Report `arXiv:1208.0806v1` [stat.ML], `arXiv.org` e-Print archive, August 2012. To download the R programs, choose "Other Formats" and then "Source".

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

[20] Vladimir Vovk. Conditional validity of inductive conformal predictors, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 5, September 2012.

A Leave-one-out conformal prediction

In this appendix we consider the extreme case where the number of folds is equal to the size of the training set, $K = l$. This special case of cross-conformal prediction will be called *leave-one-out conformal prediction*, and the corresponding predictors will be called *leave-one-out conformal predictors* (LOOCPs).

The method of leave-one-out conformal prediction is likely to have two disadvantages as compared with 5-fold or 10-fold cross-conformal prediction: first, it is computationally less efficient, and second, it may lead to loss of informational efficiency because of high variance caused by the similarity of the folds (as in the standard method of cross-validation [2, 11]). We discuss it in this appendix because of its conceptual simplicity; in particular, we will see that already in this case the analogue of Proposition 1 fails.

Let $l := 9$ and consider the exchangeable probability measure assigning the same probability $1/10!$ to each of the $10!$ permutations of a given sequence of 10 distinct examples $z_1, \dots, z_{10}$. The first 9 examples in a random permutation are assigned to the training set and the last example is the test example. Suppose the chosen inductive conformity measure A is such that

$$A((z'_1, \dots, z'_8), z'_9) \neq A((z'_1, \dots, z'_8), z'_{10})$$

for any permutation $z'_1, \dots, z'_{10}$ of $z_1, \dots, z_{10}$. We will also assume that $A(\zeta, z)$ does not depend on the ordering of $\zeta \in \mathbf{Z}^*$ for any $z \in \mathbf{Z}$, and will sometimes write $A(\zeta', z)$ for $A(\zeta, z)$ where ζ' is the (multi)set consisting of all elements of ζ. (This notation will be used only when all elements of ζ are distinct.)

With the sequence $z_1, \dots, z_{10}$ and the inductive conformity measure A we can associate the 10×10 matrix B having 1 on the main diagonal and the off-diagonal elements

$$B_{i,j} := \begin{cases} 1 & \text{if } A(\{z_1, \dots, z_{10}\} \setminus \{z_i, z_j\}, z_i) \geq A(\{z_1, \dots, z_{10}\} \setminus \{z_i, z_j\}, z_j) \\ 0 & \text{otherwise.} \end{cases}$$

The binary relation whose adjacency matrix is B is reflexive, antisymmetric, and total, but not necessarily transitive; we will identify B and this binary relation. The probability of error of Γ^ϵ, where Γ is the corresponding LOOCP and $\epsilon \in (0, 1)$, is the percentage of rows in B whose percentage of 1s is at most ϵ.

If the matrix B is transitive, it is a total order, and we can permute the

examples in such a way that it becomes

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}. \tag{13}$$

This shows that corresponding LOOCPs are unconditionally valid under our probability measure. On the other hand, it is easy to give an example of a non-transitive B that leads to a LOOCP that is not unconditionally valid: take, e.g.,

$$B := \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1
\end{pmatrix} \tag{14}$$

(reflecting a triangle in the lower left corner in the main diagonal of (13)). Corresponding LOOCPs make an error with probability 1 when $\epsilon = 0.6$. It is clear that the idea works for any value of l.

We have just seen that unconditional validity can be violated for LOOCPs, but it can be argued that in our example it is violated in a non-interesting way: in practice people are interested in small values of ϵ. To see that unconditional validity can be violated for small values of ϵ, take the analogue of (13) of size 100×100. Replacing the upper left 10×10 submatrix by (14), we obtain a probability measure and a LOOCP that makes an error with probability 10% when $\epsilon = 6\%$.

B Bootstrap conformal prediction

In this appendix we study empirically the modification of the CCP based on bootstrap ([3]; [4]; [10], Section 7.11). In the *bootstrap conformal predictor* (BCP), the following procedure is repeated K times, for $k = 1, \dots, K$. A sample $\sigma^k = (\sigma_1^k, \dots, \sigma_l^k)$ of size l (with replacement) is chosen from the index set $\{1, \dots, l\}$ of the training set $z_1, \dots, z_l$ and used to compute the conformity

Table 5: The analogue of part of Table 1 for the BCP.

Seed	0	1	...	99	Average	St. dev.
mean confidence, $K = 10$	99.17%	99.21%	...	99.23%	99.166%	0.061%
mean credibility, $K = 10$	50.96%	49.78%	...	50.74%	50.791%	0.912%

scores

$$\alpha_{i,k} := A(z_{\sigma^k}, z_i), \quad i \in \{1, \dots, l\} \setminus \Sigma^k, \qquad \alpha_k^y := A(z_{\sigma^k}, (x, y)),$$

for each potential label $y \in \mathbf{Y}$, where $z_{(\sigma_1,\dots,\sigma_l)} := (z_{\sigma_1}, \dots, z_{\sigma_l})$, Σ^k is the set of chosen indices $\Sigma^k := \{\sigma_1^k, \dots, \sigma_l^k\}$ (as Σ^k is a set, all repetitions among its elements are eliminated), and A is a given inductive conformity measure. The corresponding p-values are defined by

$$p^y := \frac{\sum_{k=1}^K \left|\left\{i \in \{1, \dots, l\} \setminus \Sigma^k \mid \alpha_{i,k} \le \alpha_k^y\right\}\right| + T/l}{T + T/l}, \tag{15}$$

where $T := \sum_{k=1}^K (l - \left|\Sigma^k\right|)$ is the total size of the *calibration sets* $z_{\{1,\dots,l\}\setminus\Sigma^k}$. (Notice that in (15) we have scaled up the constant 1 in (9) in proportion to the increase in the total size of the calibration sets from l to T. For completeness, we define (15) to be 1 in the rare cases where $T = 0$.) Confidence and credibility are now defined as usual.

Figure 6 shows that the BCP is, like the CCP, empirically well calibrated on our data set (for the analogous plots corresponding to $K = 5$ folds in this and following figures, see Online Resource 1). For results about the efficiency of the BCP see Table 5. They are not as good as for the CCP in Table 1, and comparable to the results for the ICP.

A popular modification of bootstrap is its randomized version ([4], Section 4). The randomized version of the BCP is defined similarly: the only difference from the basic version described earlier is that the labels of the examples in the *bootstrap sample* z_{σ^k} are flipped with probability 0.1 independently before computing the conformity scores. The randomized version is even less efficient than the basic version (cf. Tables 5 and 6), but it is interesting that the randomization affects not only the efficiency but also validity of the BCP: the lack of calibration in Figure 7 is obvious (although far from being as pronounced as in Figure 8 below). Figure 7 also explains the lack of efficiency of randomized BCPs as compared to basic BCPs: the former are overly conservative for small significance levels.

C An approach based on Fisher's method

In this appendix we will briefly discuss an approach to cross-conformal prediction leading to badly miscalibrated set predictions. Fisher's method [5] of

Table 6: The analogue of Table 1 for the randomized BCP.

Seed	0	1	...	99	Average	St. dev.
mean confidence, $K = 10$	98.94%	98.92%	...	98.91%	98.917%	0.073%
mean credibility, $K = 10$	51.68%	50.12%	...	50.51%	50.925%	0.820%

combining p-values $p_1, \dots, p_K$, valid when the K p-values are independent and distributed uniformly on $[0, 1]$, combines them into one statistic $-2\sum_{k=1}^{K} \ln p_k$ having the chi-squared distribution with $2K$ degrees of freedom. The corresponding p-value will be denoted $F(p_1, \dots, p_K)$:

$$F(p_1, \dots, p_K) := \mathbb{P}\left(\chi^2 \geq -2\sum_{k=1}^{K} \ln p_k\right), \tag{16}$$

where χ^2 is a random variable having the chi-squared distribution with $2K$ degrees of freedom. Even when the p-values are not distributed uniformly on $[0, 1]$ (i.e., they can be conservative, as is the case in our applications), $F(p_1, \dots, p_K)$ will still be a valid (perhaps conservative) p-value.

Naive cross-conformal predictors are defined as follows. The training set is split into K subsets, as in the case of CCPs. For each $k \in \{1, \dots, K\}$ find the p-values p_k^y via (10). Define $p^y := F(p_1^y, \dots, p_K^y)$, and then define confidence, credibility, and set predictors (1) as before. In other words, naive CCPs are defined in the same way as CCPs except that the function F is defined by (16) rather than by the expression following the $=$ in (11). Figure 8 is the analogue of the two left panels of Figure 2 for naive CCPs. It is obvious that the set predictions are very poorly calibrated; the p-values computed from different folds are heavily dependent. We do not give the efficiency results (such as those given in Table 1) for the naive CCP since efficiency without validity (at least approximate) is meaningless.

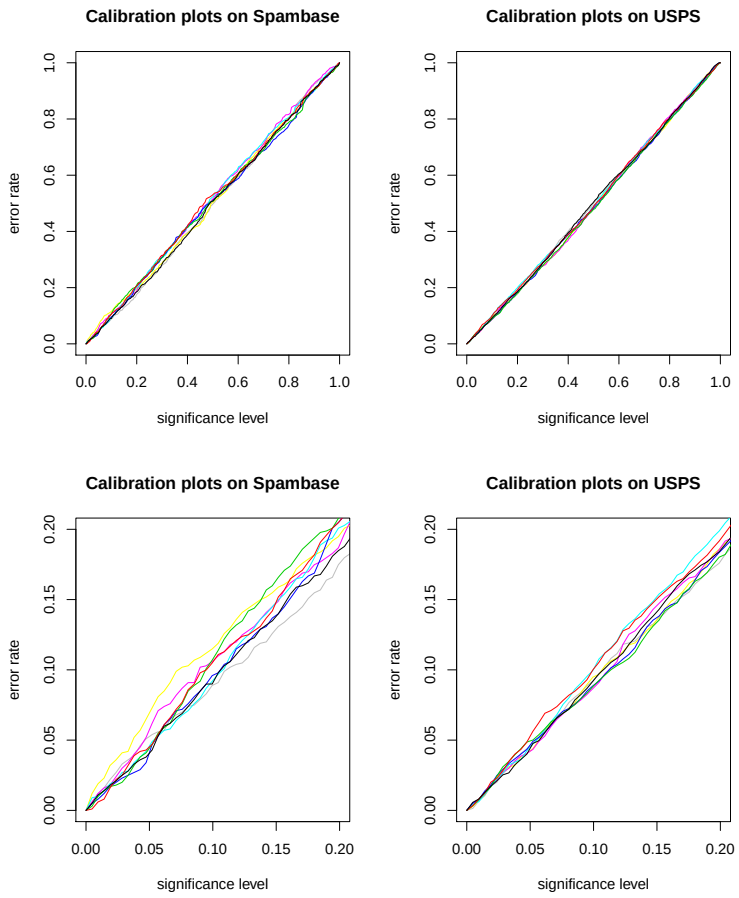

Figure 2: Top panels: the calibration plots on the Spambase (left panel) and USPS (right panel) data sets for the cross-conformal predictor with $K = 10$ folds and the first 8 seeds, 0–7, for the pseudorandom number generators. Bottom panels: the lower left corner of the corresponding top panel.

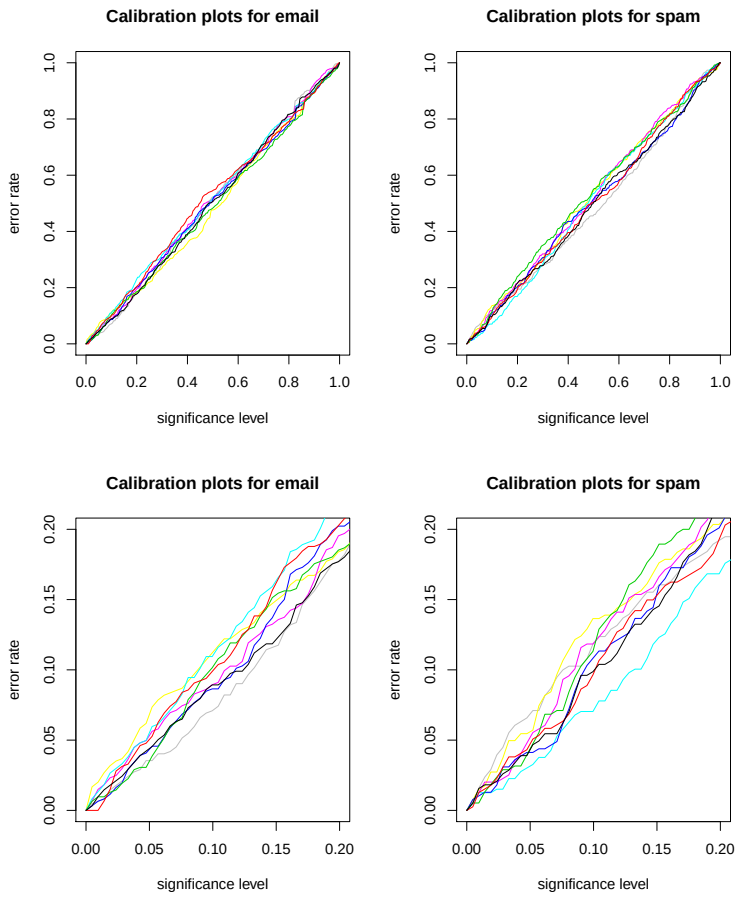

Figure 3: Top panels: the separate calibration plots for the CCCP with $K = 10$ for the examples labelled as `email` (left) and `spam` (right) in the test set and for the first 8 seeds of the pseudorandom number generator. Bottom panels: the lower left corner of the corresponding top panel.

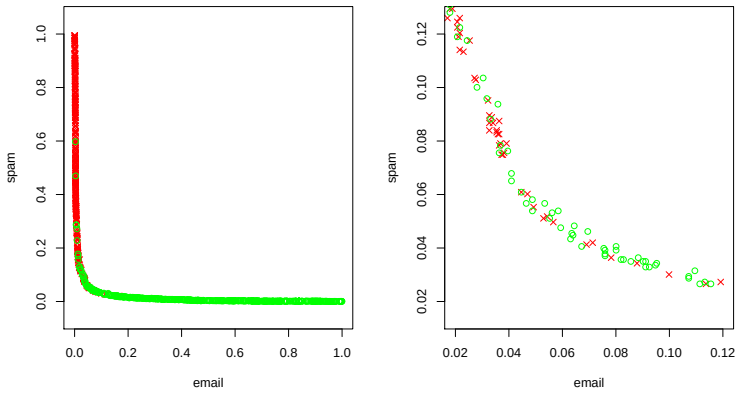

Figure 4: The scatter plot of (p^0, p^1) for the CCCP with $K = 10$ folds, all examples in the test set, and the first 8 seeds of the pseudorandom number generator. Email is shown as noughts and spam as crosses. Left panel: the full scatter plot (with spam drawn before email). Right panel: its lower left corner.

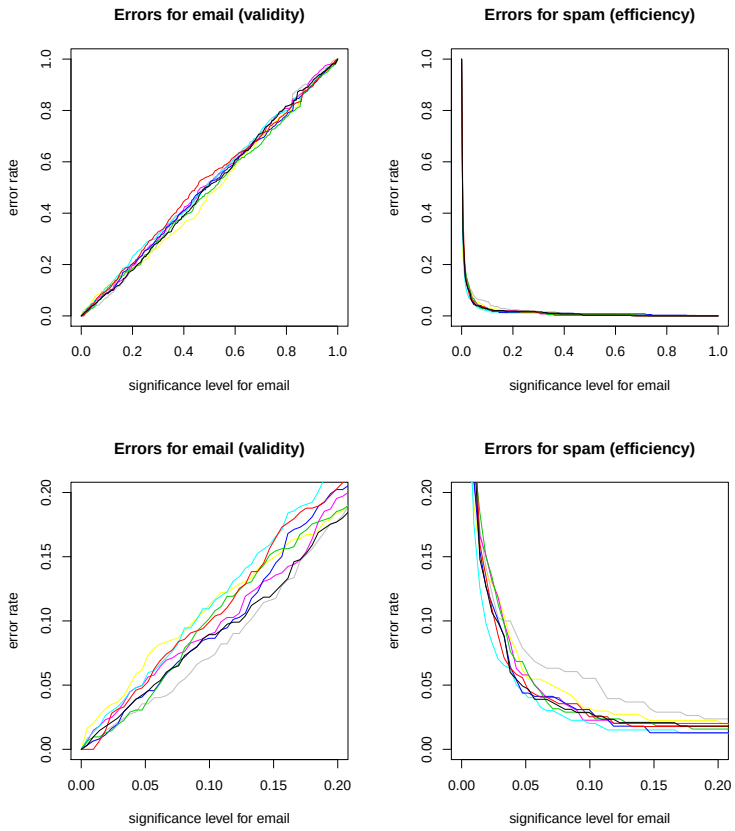

Figure 5: Top panels: the percentages of errors made by the spam filter based on the CCCP with $K = 10$ on email (left) and spam (right) for different target percentages of errors made on email and for the first 8 seeds of the pseudorandom number generator. Bottom panels: the lower left corner of the corresponding top panel.

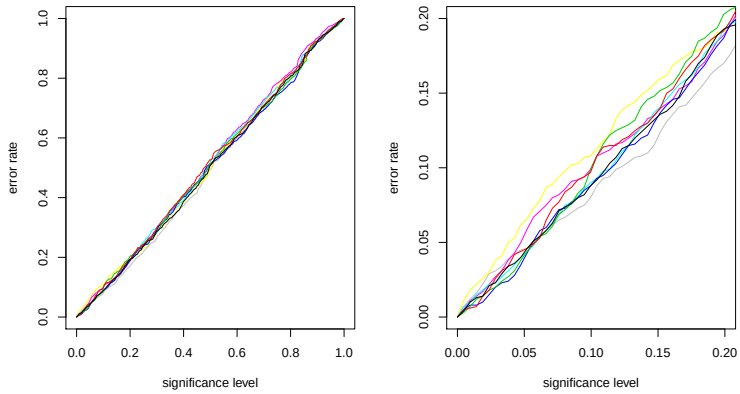

Figure 6: Left panel: the calibration plots for the 10-fold BCP, the Spambase data set, and the first 8 seeds for the pseudorandom number generator. Right panel: the lower left corner of the left panel.

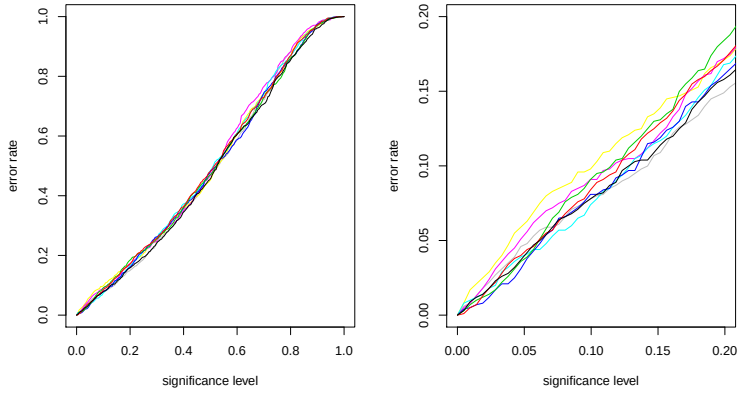

Figure 7: The analogue of Figure 6 for the 10-fold randomized BCP.

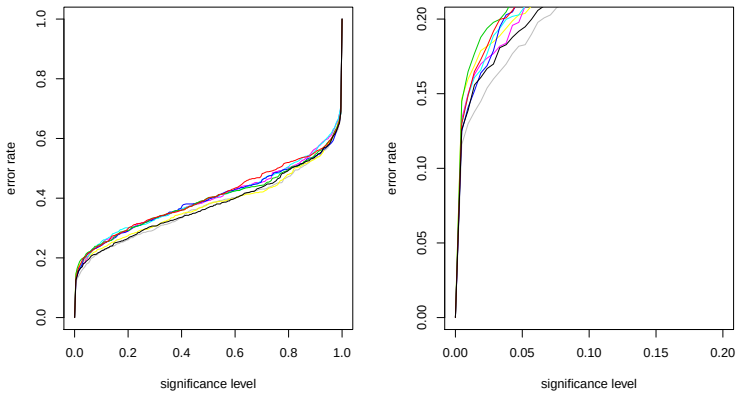

Figure 8: The analogue of Figure 6 for the 10-fold naive CCP.