INFORMATIONAL AND COMPUTATIONAL EFFICIENCY OF SET PREDICTORS

Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
United Kingdom

ABSTRACT

There are two methods of set prediction that are provably valid under the assumption of randomness: transductive conformal prediction and inductive conformal prediction. The former method is informationally efficient but often lacks computational efficiency. The latter method is, vice versa, computationally efficient but less efficient informationally. This talk discusses a new method, which we call cross-conformal prediction, that combines informational efficiency of transductive conformal prediction with computational efficiency of inductive conformal prediction. The downside of the new method is that its validity is an empirical rather than mathematical fact.

1. INTRODUCTION

The method of (transductive) conformal prediction produces set predictions that are automatically valid in the sense that their unconditional coverage probability is equal to or exceeds a preset confidence level ([1], Chapter 2). A more computationally efficient method of this kind is that of inductive conformal prediction ([2], [1], Section 4.1, [3]). 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, this talk 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. 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 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.

The empirical studies described in this paper used the R system and the `gbm` package written by Greg Ridgeway (based on the work of Freund, Schapire, and Friedman). This work was partially supported by the Cyprus Research Promotion Foundation.

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 *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 $\{y \mid p^y > \epsilon\}$.

In inductive conformal prediction 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. 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 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.

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 Section 3 we describe the method of cross-conformal prediction. Since we have no theoretical results about the validity of cross-conformal prediction in this talk, we rely on empirical studies involving the standard `Spambase` data set. Finally, we use the same data set to demonstrate the efficiency of cross-conformal predictors as compared with inductive conformal predictors. Section 4 states an open problem.

For the full version of this extended abstract, see [4].

2. 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*. 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 interested in the case where $A(\zeta, z)$ does not depend on the order of the elements of $\zeta \in \mathbf{Z}^*$). The idea behind the *conformity score* $A(z_T, z)$ is that it should measure how well the example z conforms to the proper training set z_T. A standard choice is

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

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 Section 3 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

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

and

$$\alpha_i := A(z_T, z_i), \;\; i \in C, \quad \alpha^y := A(z_T, (x, y)) \tag{3}$$

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

We call the property of inductive conformal predictors asserted in Proposition 1 unconditional validity since it is about the unconditional probability of error. Various conditional properties of validity are discussed in [5] and, in more detail, [6].

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 as the *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 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.

In our experiments reported in the next section 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. [7], 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. [7], 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 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 (CCP) 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

$$\alpha_{i,k} := A(z_{S_{-k}}, z_i), \;\; i \in S_k, \quad \alpha_k^y := A(z_{S_{-k}}, (x, y)), \tag{4}$$

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

$$p^y := \frac{\sum_{k=1}^K |\{i \in S_k \mid \alpha_{i,k} \le \alpha_k^y\}| + 1}{l + 1}. \tag{5}$$

Confidence and credibility are now defined as before; the set predictor Γ^ϵ is also defined as before, by (2), where $\epsilon > 0$ is another parameter.

The definition of CCPs parallels that of ICPs, except that now we use the whole training set for calibration. The conformity scores (4) are computed as in (3) but using the union of all the folds except for the current one as the proper training set. Calibration (5) 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

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

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, $|S_1| = \dots = |S_K|$, a simple calculation gives

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

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 $K \ll l$.

We give calibration plots for 5-fold and 10-fold cross-conformal prediction taking $K \in \{5, 10\}$ following the advice in [7] (who refer to Breiman and Spector's and Kohavi's work). In our experiments we use the popular `Spambase` data set. The size of the data set is 4601, and there are two labels: `spam`, encoded as 1, and `email`, encoded as 0.

We consider the conformity measure (1) where f is output by MART ([7], Chapter 10) and

$$\Delta(y, f(x)) := \begin{cases} f(x) & \text{if } y = 1 \\ -f(x) & \text{if } y = 0. \end{cases} \tag{6}$$

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 (6) as conformity score very natural. (MART is known [7] to give good results on the `Spambase` dataset.)

Figure 1 gives the calibration plots for the CCP and for 8 random splits of the data set into a training set of size 3600 and a test set of size 1001 and of the training set into 5 or 10 folds. There is a further source of randomness as the MART procedure is itself randomized. The functions plotted in Figure 1 map each significance level ϵ to the percentage of erroneous predictions made by the set predictor Γ^ϵ on the test set. Visually, the plots are well-calibrated (close to the bisector of the first quadrant).

As for the efficiency of the CCP, see Table 1, which gives some statistics for the confidence and credibility output by the ICP and the 5-fold and 10-fold CCP. The columns labelled "0" to "7" give the mean values of confidence and credibility over the test set for various values of the seed for the R pseudorandom number generator. The column labelled "Average" gives the average

$$\bar{v} := \frac{1}{8}\sum_{i=0}^7 v_i$$

of all the 8 mean values (which we denote $v_0, \dots, v_7$) for the seeds 0–7, and the column labelled "St. dev." gives the standard unbiased estimate

$$\sqrt{\frac{1}{7}\sum_{i=0}^7 (v_i - \bar{v})^2}$$

of the standard deviation of the mean values computed from $v_0, \dots, v_7$. The biggest advantage of the CCP is in the stability of its confidence values: the standard deviation of the mean confidences is much less than that for the ICP. However, the CCP also gives higher confidence; to some degree this can be seen from the table, but the high variance of the ICP confidence masks it: e.g., for the first 100 seeds the average of the mean confidence for ICP is 99.16% (with the standard deviation of the mean confidences equal to 0.149%, corresponding to the standard deviation of 0.015% of the average mean confidence).

4. CONCLUSION

At this time there are no theoretical results about the validity of cross-conformal predictors (like Proposition 1), and it is an interesting open problem to establish such results.

5. REFERENCES

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

[2] 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)*, Las Vegas, NV, 2002, pp. 159–163, CSREA Press.

[3] Anonymous, "Generalized conformal prediction for functional data," Submitted to NIPS 2012, June 2012.

[4] Vladimir Vovk, "Cross-conformal predictors," Tech. Rep. `arXiv:1208.0806v1` [stat.ML], `arXiv.org` e-Print archive, August 2012.

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

[6] Anonymous, "Inductive conformal predictors in the batch mode," Submitted to ACML 2012, July 2012.

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

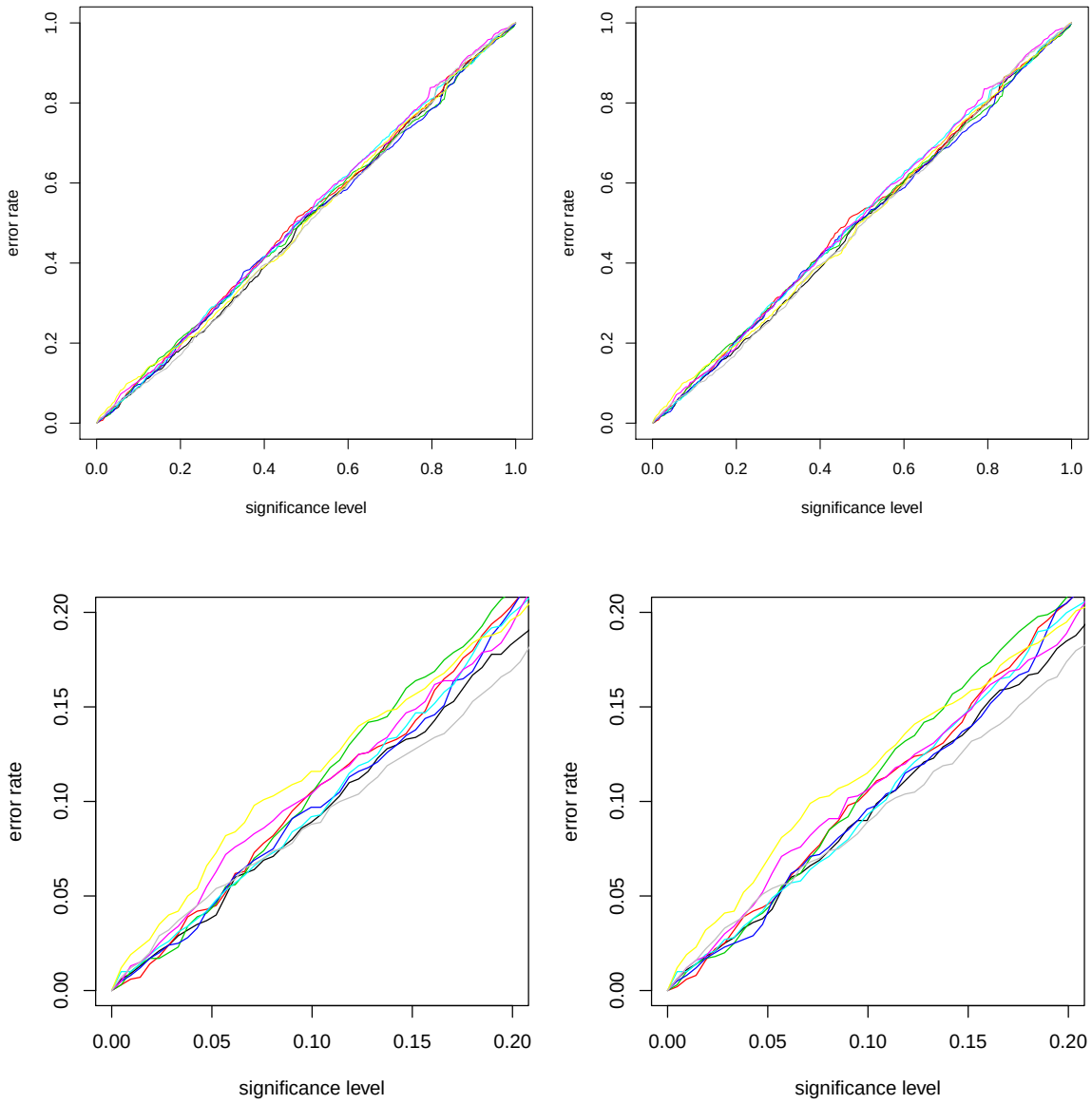

Figure 1. Top panels: the calibration plots for the cross-conformal predictor with $K = 5$ (left) and $K = 10$ (right) folds and the first 8 seeds, 0–7, for the R pseudorandom number generator. Bottom panels: the lower left corner of the corresponding top panel (which is the most important part of the calibration plot in applications).

Seed	0	1	2	3	4	5	6	7	Average	St. dev.
mean conf., ICP	99.25%	99.23%	99.00%	99.17%	99.30%	99.12%	99.38%	99.25%	99.21%	0.116%
mean cred., ICP	51.31%	50.37%	49.93%	52.45%	48.98%	50.34%	50.18%	52.00%	50.69%	1.148%
mean conf., $K = 5$	99.22%	99.17%	99.17%	99.24%	99.27%	99.27%	99.30%	99.30%	99.24%	0.054%
mean cred., $K = 5$	51.11%	49.74%	50.34%	50.69%	49.85%	49.49%	50.95%	51.46%	50.45%	0.713%
mean conf., $K = 10$	99.24%	99.20%	99.20%	99.23%	99.26%	99.28%	99.34%	99.32%	99.26%	0.051%
mean cred., $K = 10$	51.08%	49.74%	50.29%	50.77%	49.75%	49.48%	50.96%	51.45%	50.44%	0.727%

Table 1. Mean (over the test set) confidence and credibility for the ICP and the 5-fold and 10-fold CCP. The results are given for various values of the seed for the R pseudorandom number generator; column "Average" gives the average of all the 8 values for the seeds 0–7, and column "St. dev." gives the standard unbiased estimate of the standard deviation computed from those 8 values.