Cross-conformal e-prediction

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 #26

January 16, 2020

Project web site:
http://alrw.net

Abstract

This note discusses a simple modification of cross-conformal prediction inspired by recent work on e-values. The precursor of conformal prediction developed in the 1990s by Gammerman, Vapnik, and Vovk was also based on e-values and is called *conformal e-prediction* in this note. Replacing e-values by p-values led to conformal prediction, which has important advantages over conformal e-prediction without obvious disadvantages. The situation with cross-conformal prediction is, however, different: whereas for cross-conformal prediction validity is only an empirical fact (and can be broken with excessive randomization), this note draws the reader's attention to the obvious fact that cross-conformal e-prediction enjoys a guaranteed property of validity.

Contents

1 Introduction

Conformal prediction is based on the notion of a p-value. At this time p-values are widely discussed (see, e.g., [1]), and several alternatives to p-values have been proposed. Perhaps the most popular alternative is Bayes factors, which, when stripped of their Bayesian context, are referred to as e-values in [10]. In fact, e-values were used (under the name of i-values) when discussing precursors of conformal prediction in the 1990s. One early description is [3]. In this note, it will be convenient to distinguish between *conformal e-prediction* (using e-values) and *conformal p-prediction* (standard conformal prediction using p-values).

Conformal e-prediction was superseded by conformal (p-)prediction mainly for two reasons:

1. Conformal predictions can be packaged as prediction sets [9, Section 2.2], in which case the property of validity of conformal predictors is very easy to state: we just say that the probability of error is at most ϵ for a pre-specified significance level ϵ [9, Proposition 2.3].

2. In the on-line mode of prediction, smoothed conformal predictors make errors independently, and so small probabilities of errors manifest themselves, with high probability, as a low frequency of errors [9, Corollary 2.5].

These are important advantages of conformal p-prediction, which do not appear to be counterbalanced by any clear disadvantages.

Whereas the notion of a conformal e-predictor does not appear particularly useful, using e-values in place of p-values in cross-conformal prediction [8] has a clear advantage. Cross-conformal predictors are not provably valid [8, Appendix], and this sometimes even shows in experimental results [6]. The limits of violations of validity are given by Rüschendorf's result (see, e.g., [11]): when merging p-values coming from different folds by taking arithmetic mean (this is essentially what cross-conformal predictors do), the resulting arithmetic mean has to be multiplied by 2 in order to guarantee validity. In the recent method of jackknife+, introduced in [2] and closely related to cross-conformal prediction, there is a similar factor of 2 [2, Theorem 1], which cannot be removed in general [2, Theorem 2].

The situation with e-values is different: the arithmetic mean of e-values is always an e-value. This is an obvious fact, but it is shown in [10] that arithmetic mean is the only useful merging rule. Therefore, the version of cross-conformal prediction based on e-values, which we call *cross-conformal e-prediction* in this note, is always valid.

2 Conformal e-predictors

Suppose we are given a training set $z_1, \ldots, z_n$ consisting of labelled objects $z_i = (x_i, y_i)$ and our goal is to predict the label of a new object x. In this note we consider predictors of the following type: for each potential label y for x we

would like to have a number $f(z_1, \ldots, z_n, x, y)$ reflecting the plausibility of y being the true label of x. An example is conformal transducers [9, Section 2.5], which, in the terminology of this note, may be called *conformal p-predictors*. The output

$$y \mapsto f(z_1, \ldots, z_n, x, y)$$

of a conformal p-predictor is the full conformal prediction for the label of x; e.g., it determines the prediction set at each significance level. We will sometimes write $f(z_1, \ldots, z_n, z)$, where $z := (x, y)$, instead of $f(z_1, \ldots, z_n, x, y)$.

We will use the notation $\mathbf{X}$ for the object space and $\mathbf{Y}$ for the label space (both assumed non-empty). These are measurable spaces from which the objects and labels, respectively, are drawn. Full observations $z = (x, y)$ are drawn from $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. For any non-empty set X, X^+ will be the set $\cup_{n=1}^{\infty} X^n$ of all non-empty finite sequences of elements of X.

A *conformal e-predictor* is a function $f : \mathbf{Z}^+ \to [0, \infty)^+$ that maps any finite sequence $(z_1, \ldots, z_m)$, $m \in \{1, 2, \ldots\}$, to a finite sequence $(\alpha_1, \ldots, \alpha_m)$ of nonnegative numbers with average at most 1,

$$\frac{1}{m} \sum_{i=1}^{m} \alpha_i \leq 1, \tag{1}$$

that satisfies the following property of equivariance: for any $m \in \{2, 3, \ldots\}$, any permutation π of $\{1, \ldots, m\}$, any $(z_1, \ldots, z_m) \in \mathbf{Z}^m$, and any $(\alpha_1, \ldots, \alpha_m) \in [0, \infty)^m$,

$$(\alpha_1, \ldots, \alpha_m) = f(z_1, \ldots, z_m) \Longrightarrow (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(m)}) = f(z_{\pi(1)}, \ldots, z_{\pi(m)}).$$

The conformal e-predictor proposed in [4] is

$$f(z_1, \ldots, z_m) := \begin{cases} m/\, |\text{SV}| & \text{if } m \in \text{SV} \\ 0 & \text{otherwise,} \end{cases}$$

where SV is the set of indices of support vectors: $i \in \text{SV}$ if and only if z_i is a support vector for the SVM constructed from $z_1, \ldots, z_m$ as training set. When given a training set $z_1, \ldots, z_n$ and a new object x, this conformal e-predictor goes through all potential labels y for x and for each constructs an SVM and outputs $f(z_1, \ldots, z_n, x, y)$. It makes it computationally inefficient.

The following obvious proposition asserts the validity of conformal e-predictors.

Proposition 1. *For any n, if $z_1, \ldots, z_n, (x, y)$ are IID,*

$$\mathbb{E} f(z_1, \ldots, z_n, x, y) \leq 1.$$

3 Split conformal e-predictors

Let us fix a measurable space Σ (a *summary space*). A Σ-*valued split conformity measure* is a measurable function $A : \mathbf{Z}^+ \to \Sigma$. Intuitively, $A(z_1, \ldots, z_m, z)$

encodes how well z conforms to $z_1, \ldots, z_m$. A *normalizing transformation* N : $\Sigma^+ \to [0, \infty)^+$ is an equivariant measurable function that maps every non-empty finite sequence $(\sigma_1, \ldots, \sigma_m)$ of elements of Σ to a finite sequence $(\alpha_1, \ldots, \alpha_m)$ of the same length of nonnegative numbers whose average is at most 1 (i.e., satisfying (1)).

To apply split conformal e-prediction to a training set $z_1, \ldots, z_n$, we split it into two parts, the *training set proper* $z_1, \ldots, z_{n-c}$ and the *calibration set* $z_{n-c+1}, \ldots, z_n$. For a new object x and a potential label y for it, we set

$$f(z_1, \ldots, z_n, x, y) := \alpha^y \tag{2}$$

where α^y is defined using the following steps:

$$\sigma_i := A(z_1, \ldots, z_{n-c}, z_{n-c+i}), \quad i = 1, \ldots, c,$$
$$\sigma^y := A(z_1, \ldots, z_{n-c}, (x, y)),$$
$$(\alpha_1^y, \ldots, \alpha_c^y, \alpha^y) := N(\sigma_1, \ldots, \sigma_c, \sigma^y).$$

For many choices of A and N, the split conformal e-predictor (2) will be computationally efficient; this is the case when:

1. Processing the training set proper only once, we can find an easily computable rule transforming z into $A(z_1, \ldots, z_{n-c}, z)$.

2. The normalizing transformation N is easily computable.

An example of an easily computable normalizing transformation is

$$(\sigma_1, \ldots, \sigma_m) \mapsto \frac{1}{\sum_{i=1}^m \sigma_i} (\sigma_1, \ldots, \sigma_m),$$

where the summary space is supposed to be $(0, \infty)$.

Proposition 1, our statement of validity, continues to hold for split conformal e-predictors.

4 Cross-conformal e-predictors

A Σ-valued split conformity measure A is a Σ-*valued cross-conformity measure* if $A(z_1, \ldots, z_m, z)$ does not depend on the order of its first m arguments. Given such an A and a normalizing transformation N, the corresponding *cross-conformal e-predictor* (CCEP) is defined as follows. The training sequence $z_1, \ldots, z_n$ is randomly split into K non-empty multisets (*folds*) z_{S_k}, $k = 1, \ldots, K$, of equal (or as equal as possible) sizes $|S_k|$, where $K \in \{2, 3, \ldots\}$ is a parameter of the algorithm, $(S_1, \ldots, S_K)$ is a partition of the index set $\{1, \ldots, n\}$, and z_{S_k} consists of all z_i, $i \in S_k$. For each $k \in \{1, \ldots, K\}$ and each potential label $y \in \mathbf{Y}$ of the new object x, find the output α_k of the split conformal e-predictor on the new object x and its postulated label y with $z_{S_{-k}}$ as training set proper and z_{S_k} as calibration set, where $S_{-k} := \cup_{j \neq k} S_j = \{1, \ldots, n\} \backslash S_k$

is the complement to the fold S_k. The corresponding CCEP is defined by

$$f(z_1, \ldots, z_n, (x, y)) := \frac{1}{K} \sum_{k=1}^{K} \alpha_k.$$

(A slight modification, still provably valid, of this definition is where the arithmetic mean is replaced by the weighted mean with the weights proportional to the sizes $|S_k|$ of the folds.)

Proposition 1 still holds for cross-conformal e-predictors; this trivially follows from the arithmetic mean of e-values being an e-value.

Remark 2. To compare the outputs of cross-conformal p-predictors (CCPP) and CCEP, we can use the rough transformation discussed in [10]: a p-value of p roughly corresponds to an e-value of $1/p$. Under this transformation, the arithmetic average of e-values corresponds to the harmonic average of p-values, and the harmonic average is always less than or equal to the arithmetic average [5, Theorem 16]. This suggests that CCEP produce better results than CCPP do. In the opposite direction, the arithmetic average of p-values corresponds to the harmonic average of e-values, which again suggests that CCEP produce a better result than CCPP do.

5 Validity in the time domain

Advantage 1 of conformal p-prediction given on p. 1 is that, when discussing validity, we can talk about probabilities instead of p-values. It mostly disappears when we move to conformal e-prediction: validity has to be defined in terms of e-values (however, it has been argued [7] that e-values are more intuitive than p-values). In this section we discuss advantage 2, which requires the online mode of prediction. We will see that it still holds, albeit in a weakened form.

The notion of validity asserted in Proposition 1 (applied to CCEP) is stated in the "space domain": the output of the CCEP on the true label is an e-value, i.e., its mean value at a fixed time over the probability space does not exceed 1. Now we will complement the validity in the space domain by the validity in the time domain assuming that the CCEP is bounded.

In the online prediction protocol, we observe an object x_1, apply the CCEP to compute the values $f(x_1, y)$ for all possible labels $y \in \mathbf{Y}$, observe the true label y_1, observe another object x_2, apply the CCEP to compute the values $f(x_1, y_1, x_2, y)$ for all possible labels $y \in \mathbf{Y}$, observe the true label y_2, etc.

Remark 3. Intuitively, using CCEP in the online mode defeats the purpose of cross-conformal prediction: once we process $x_1, y_1, \ldots, x_n, y_n$, we would like to apply the rules that we have found (see item 1 on p. 3) to a large number of new objects rather than just one. But more realistic settings are also more complicated.

Let the values for the true labels be $e_n := f(x_1, y_1, \ldots, x_n, y_n)$. It follows

4

from, e.g., [9, Lemma 3.15] that, if the observations (x_i, y_i) are IID,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} e_i \leq 1.$$

We can see that the long-term time average of the values for the true labels is bounded above by 1. In this sense they are time-wise e-values.

Acknowledgments

References

[1] Special issue on p-values. *American Statistician*, 73(Supplement 1), 2019.

[2] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. Technical Report arXiv:1905.02928 [stat.ME], arXiv.org e-Print archive, December 2019.

[3] Alex Gammerman, Vladimir Vapnik, and Vladimir Vovk. Transduction in pattern recognition, 1997. Manuscript submitted to the Fifteenth International Joint Conference on Artificial Intelligence in January 1997. Extended version published as [4].

[4] Alex Gammerman, Vladimir Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155, San Francisco, CA, 1998. Morgan Kaufmann.

[5] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, second edition, 1952.

[6] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. *Proceedings of Machine Learning Research*, 60:154–173, 2017.

[7] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper 54, October 2019 (first posted March 2019).

[8] Vladimir Vovk. Cross-conformal predictors, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 6, January 2013 (first posted August 2012).

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

[10] Vladimir Vovk and Ruodu Wang. Combining e-values and p-values. Technical Report arXiv:1912.06116 [math.ST], arXiv.org e-Print archive, December 2019.

[11] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging, Online Compression Modelling project (New Series), `http://alrw.net`, Working Paper 21, November 2019 (first posted November 2017).