Transductive 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 #8

First posted April 24, 2013. Last revised October 13, 2015.

Project web site:
http://alrw.net

Abstract

This paper discusses transductive versions of conformal predictors and inductive conformal predictors. The transductive versions are computationally inefficient for long test sequences, but it turns out that apparently crude "Bonferroni predictors" are about as good in their informational efficiency and vastly superior in computational efficiency. The paper explores transductive predictors both theoretically and experimentally, in the latter case using the standard USPS data set of handwritten digits.

Contents

1 Introduction

The most standard learning problems are inductive: given a training sequence of labelled objects, the task is to come up with a prediction rule with a reasonable performance on unknown test objects. In typical transductive problems (Vapnik and Chervonenkis[10], Chapter VI, Sections 10–13; Vapnik[9], Chapter 8) we are given both a training sequence of labelled objects and a test sequence of unlabelled objects; the task is to come up with a prediction rule, which may depend on both sequences, with a reasonable performance on the test sequence.

Conformal prediction (see, e.g., Vovk et al.[13]) is a set of methods for producing prediction regions rather than point predictions. Typical predictors of this kind are conformal predictors (Vovk et al.[13], Chapter 2), inductive conformal predictors (Vovk et al.[13], Section 4.1), and cross-conformal predictors (Vovk[14]). None of these methods is transductive in the sense of Vapnik and Chervonenkis[10] and Vapnik[9] (although conformal predictors do have a transductive flavour, as discussed in Section 7 below).

The goal of this paper is to introduce fully transductive versions of conformal and related predictors. The basic definitions are given in Section 2. Section 3 introduces Bonferroni predictors, a simple and computationally efficient modification of conformal predictors adapted to the transductive framework. Sections 4 and 5 contain simple theoretical results about transductive conformal predictors and Bonferroni predictors. Section 6 reports on experimental results. Section 7 discusses the notion of transduction as used in this paper and in existing literature. Section 8 concludes the main part of the paper. A applies the same ideas to inductive conformal predictors, and B discusses disadvantages of the standard statistical approach to combining ranks in the context of transductive conformal prediction.

The expression "transductive conformal predictors" has been used before (see, e.g., Nouretdinov et al.[5]) to refer to what is called "conformal predictors" in Vovk et al.[13] and this paper. This usage agrees with the terminology of this paper, since conformal predictors are a special case of our transductive conformal predictors corresponding to a test sequence of length 1.

2 Transductive Conformal Predictors

Let $z_1 = (x_1, y_1), \dots, z_l = (x_l, y_l)$ be a training sequence and $x_{l+1}, \dots, x_{l+k}$ be a test sequence. The test sequence is a finite sequence of *objects* $x_j \in \mathbf{X}$ and the training sequence is a finite sequence of labelled objects, or *observations*, $z_i = (x_i, y_i) \in \mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. The *object space* $\mathbf{X}$, *label space* $\mathbf{Y}$, and *observation space* $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ are fixed throughout the paper; they are assumed to be measurable spaces. The set of all finite sequences of elements of $\mathbf{Z}$ is denoted $\mathbf{Z}^*$; similar notation will be used for other sets as well (such as $\mathbb{R}^*$ for the set of all finite sequences of real numbers).

Transductive conformal predictors are determined by their transductive nonconformity measures, which are defined as follows. A *transductive nonconfor-*

mity measure is a measurable function $A : \mathbf{Z}^* \times \mathbf{Z}^* \to \mathbb{R}$ such that $A(\zeta_1, \zeta_2)$ does not depend on the ordering of ζ_1. (For the specific transductive nonconformity measures used in this paper $A(\zeta_1, \zeta_2)$ will not depend on the ordering of ζ_2 either.) The intuition is that $A(\zeta_1, \zeta_2)$ (the *transductive nonconformity score*) measures the lack of conformity of the "test sequence" ζ_2 to the "training sequence" ζ_1.

The *transductive conformal predictor* (*TCP*) corresponding to A finds the prediction region for the test sequence $x_{l+1}, \dots, x_{l+k}$ at a *significance level* $\epsilon \in (0, 1)$ as follows:

- For each possible sequence of labels $(\upsilon_1, \dots, \upsilon_k) \in \mathbf{Y}^k$:
  - set $y_j := \upsilon_{j-l}$ and $z_j := (x_j, y_j)$ for $j = l+1, \dots, l+k$;
  - compute the transductive nonconformity scores

    $$\alpha_S := A(z_{\{1,\dots,l+k\}\setminus S}, z_S),$$

    where S ranges over all $(l+k)!/l!$ ordered subsets $(s_1, \dots, s_k)$ of size k of the set $\{1, \dots, l+k\}$, z_S stands for the sequence $(z_{s_1}, \dots, z_{s_k})$ (when $S = (s_1, \dots, s_k)$), and $z_{\{1,\dots,l+k\}\setminus S}$ stands for z_B, B being any ordering of $\{1, \dots, l+k\} \setminus S'$ and S' being the set of all elements of S (it does not matter which ordering is chosen, by the definition of a transductive nonconformity measure);
  - compute the p-value

    $$p(\upsilon_1, \dots, \upsilon_k) := \frac{\left|\{S \mid \alpha_S \geq \alpha_{(l+1,\dots,l+k)}\}\right|}{(l+k)!/l!}, \tag{1}$$

    where S ranges, as before, over all $(l+k)!/l!$ ordered subsets of $\{1, \dots, l+k\}$ of size k, and $|\cdots|$ stands for the size of a set.
- Output the prediction region

$$\Gamma^\epsilon(z_1, \dots, z_l, x_{l+1}, \dots, x_{l+k}) := \left\{(\upsilon_1, \dots, \upsilon_k) \in \mathbf{Y}^k \mid p(\upsilon_1, \dots, \upsilon_k) > \epsilon\right\}. \tag{2}$$

Smoothed TCPs are defined in the same way except that (1) is replaced by

$$p(\upsilon_1, \dots, \upsilon_k) := \frac{\left|\{S \mid \alpha_S > \alpha_{(l+1,\dots,l+k)}\}\right| + \theta\left|\{S \mid \alpha_S = \alpha_{(l+1,\dots,l+k)}\}\right|}{(l+k)!/l!},$$

where θ are random variables distributed uniformly on $[0, 1]$ (no independence between different sequences of postulated labels $\upsilon_1, \dots, \upsilon_k$ is required, but later on when we consider the online prediction protocol we will assume that θ are independent between different trials).

A *nonconformity measure* can now be defined as the restriction of a transductive nonconformity measure to the domain $\mathbf{Z}^* \times \mathbf{Z}$ (we identify a 1-element sequence with its only element). Nonconformity measures are well studied and

there are many useful examples of them (see, e.g., Vovk et al.[13]). For example, a natural choice of a nonconformity measure is

$$A(\zeta, (x, y)) := \Delta(y, f(x)) \,, \tag{3}$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is a prediction rule found from ζ as the training sequence and $\Delta : \mathbf{Y} \times \mathbf{Y}' \to \mathbb{R}$ is a distance between a label and a prediction. (Usually $\mathbf{Y}' \supseteq \mathbf{Y}$, such as $\mathbf{Y}' = [0, 1] \supseteq \{0, 1\} = \mathbf{Y}$.)

An interesting class of transductive nonconformity measures can be obtained from nonconformity measures. Let $\mathbb{R}$ be the set of real numbers. A *simple nonconformity aggregator* is a function $M : \mathbb{R}^* \to \mathbb{R}$ that is symmetric and increasing in each argument. (The requirement that M be symmetric, i.e., $M(\zeta)$ not depend on the ordering of ζ, is not necessary but convenient for the following discussion. The requirement that M be increasing in each argument is not necessary either but very natural.) With each nonconformity measure A and simple nonconformity aggregator M we can associate the transductive nonconformity measure

$$A_M((z_1, \dots, z_l), (z_{l+1}, \dots, z_{l+k})) := M(\alpha_{l+1}, \dots, \alpha_{l+k}) \,,$$

where

$$\alpha_j := A((z_1, \dots, z_l, z_{l+1}, \dots, z_{j-1}, z_{j+1}, \dots, z_{l+k}), z_j), \quad j = l+1, \dots, l+k \,. \tag{4}$$

Our experiments in Section 6 use the *Nearest Neighbour nonconformity measure*

$$A\big(((x_1, y_1), \dots, (x_n, y_n)), (x, y)\big) := \frac{\min_{i=1,\dots,n:y_i=y} d(x, x_i)}{\min_{i=1,\dots,n:y_i \neq y} d(x, x_i)} \,, \tag{5}$$

where d is a distance, and the *max nonconformity aggregator*

$$M(\alpha_1, \dots, \alpha_k) := \max(\alpha_1, \dots, \alpha_k) \,. \tag{6}$$

Remark. Alternatively, we could set $\alpha_j := A((z_1, \dots, z_l), z_j)$ in (4) (cf. (13) below), but this would adversely affect the already low computational efficiency of TCPs in our experiments in Section 6.

Rank-based Transductive Conformal Predictors

The notion of a simple nonconformity aggregator can be generalized as follows. A *nonconformity aggregator* is a function $M : \mathbb{R}^* \times \mathbb{R}^* \to \mathbb{R}$ such that $M(\zeta_1, \zeta_2)$ depends neither on the ordering of ζ_1 nor on the ordering on ζ_2. (The most natural class of nonconformity aggregators is where $M(\zeta_1, \zeta_2)$ is decreasing in every element of ζ_1 and increasing in every element of ζ_2, but it is too narrow for our purposes.) With each nonconformity measure A and nonconformity aggregator M we associate the transductive nonconformity measure

$$A_M((z_1, \dots, z_l), (z_{l+1}, \dots, z_{l+k})) := M\big((\alpha_1, \dots, \alpha_l), (\alpha_{l+1}, \dots, \alpha_{l+k})\big)$$

where

$$\alpha_i := A((z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_l, z_{l+1}, \ldots, z_{l+k}), z_i), \qquad i = 1, \ldots, l, \tag{7}$$

and $\alpha_{l+1}, \ldots, \alpha_{l+k}$ are defined by (4). We identify each simple nonconformity aggregator M with the nonconformity aggregator

$$M^\dagger((\alpha_1, \ldots, \alpha_l), (\alpha_{l+1}, \ldots, \alpha_{l+k})) := M(\alpha_{l+1}, \ldots, \alpha_{l+k})\,.$$

For transductive nonconformity measures obtained from nonconformity measures and nonconformity aggregators, the p-value (1) as function of the nonconformity scores α_i of individual observations reduces to the well-known notion of a one-sided permutation test (see, e.g., Lehmann[4], Section 1.7.E). In classical nonparametric statistics, the most popular permutation tests are rank tests, and we will give corresponding definitions in our current context. Let $\mathbb{N} := \{1, 2, \ldots\}$. A (simple) *rank aggregator* is a function $M : \mathbb{N}^* \to \mathbb{N}$ that is symmetric and increasing in each argument. The corresponding nonconformity aggregator is

$$M'\big((\alpha_1, \ldots, \alpha_l), (\alpha_{l+1}, \ldots, \alpha_{l+k})\big) := M(R_{l+1}, \ldots, R_{l+k})\,, \tag{8}$$

where $R_1, \ldots, R_{l+k}$ are the ranks of $\alpha_1, \ldots, \alpha_{l+k}$, respectively, in the multiset $\wr\alpha_1, \ldots, \alpha_{l+k}\wr$. Formally, R_i is defined as

$$R_i := |\{j = 1, \ldots, l+k \mid \alpha_j < \alpha_i\}| + 1\,.$$

If there are no ties (i.e., equal elements in $\wr\alpha_1, \ldots, \alpha_{l+k}\wr$), this is the usual notion of a rank; in the presence of ties, our definition is somewhat non-standard giving each tie the smallest of the ranks that it spans. (And this definition causes a counterintuitive behaviour of the definition (8), where M' is not necessarily increasing in α_j, $j \in \{l+1, \ldots, l+k\}$, even in the case where M is the max nonconformity aggregator (6).) (In classical nonparametric statistics, R_i is usually defined as

$$R_i := \frac{|\{j = 1, \ldots, n \mid \alpha_j \le \alpha_i\}| + |\{j = 1, \ldots, n \mid \alpha_j < \alpha_i\}| + 1}{2}\,,$$

where $n := l + k$. Informally, the ranks are computed as follows: rank the α_i starting from 1 and give each tie a rank equal to the average of the ranks it spans. Such R_i are also known as *midranks*: cf. Lehmann[4], Section 1.4.)

The most popular rank aggregator in classical nonparametric statistics is the *ranksum aggregator*

$$M(R_1, \ldots, R_k) := R_1 + \cdots + R_k\,, \tag{9}$$

which is used in the Wilcoxon ranksum test (see Wilcoxon[15] or Lehmann[4], Section 1.2). Using the ranksum aggregator, however, produces very poor results (see B) when the efficiency of TCPs is measured by the number of multiple predictions that they produce, as in this paper (see Section 6 below).

Remark. The classical Wilcoxon ranksum test as applied to nonconformity scores consists in computing the p-value (1) from the nonconformity aggregator corresponding to the ranksum aggregator (9). More generally, each rank aggregator M defines a statistical test for testing whether two independent samples are coming from the same distribution (with sum corresponding to the classical one-sided Wilcoxon ranksum test[15, 4]). Let S_1 and S_2 be two finite sets of real numbers; suppose, for simplicity, that all their elements are different. Find the ranks of S_2 in the merged set $S_1 \cup S_2$. Apply M to those ranks; let the result be t. The p-value $p_M(S_1, S_2)$ produced by the test (the *M-test*) is equal to the probability that the value of M will be at least t under the null hypothesis. In other words, $p_M(S_1, S_2)$ is equal to the probability that the value of M on a random sample of size $|S_2|$ without replacement from $\{1, \dots, |S_2| + |S_2|\}$ will be at least t.

Notice that the nonconformity aggregator (6) is equivalent (in the sense of leading to the same TCP) to the rank aggregator $M'(R_1, \dots, R_k) := \max(R_1, \dots, R_k)$. The corresponding TCP will be called the *rankmax TCP* (and the TCP corresponding to (9) will be called the *ranksum TCP*).

It is easy to give an explicit representation of the rankmax TCP. Remember that the length of the training sequence is l and the length of the test sequence is k and suppose that the value of the *rankmax test statistic* $\max(R_{l+1}, \dots, R_{l+k})$ is t. The probability that a random subset $\{s_1, \dots, s_k\}$ of $\{1, \dots, l+k\}$ of size k will lead to a value of the test statistic $\max(R_{s_1}, \dots, R_{s_k})$ of at least t can be found as 1 minus the probability that a random subset of $\{1, \dots, l+k\}$ of size k is covered by a fixed subset of $\{1, \dots, l+k\}$ of size $t-1$ (namely, by the set of indices i with $R_i < t$). In other words, the p-value is

$$1 - \frac{\binom{t-1}{k}}{\binom{l+k}{k}} = 1 - \frac{(t-1)!\, l!}{(t-1-k)!\, (l+k)!} \tag{10}$$

(which is understood to be 1 when $t \le k$).

Remark. The smallest possible value of (10), attained when $t = l + k$, is

$$1 - \frac{\binom{t-1}{k}}{\binom{l+k}{k}} = \frac{k}{l+k}\,.$$

3 Bonferroni Predictors

Unfortunately, transductive conformal predictors are computationally inefficient, especially if we want to predict many test objects at once: we have to go over all $|\mathbf{Y}|^k$ combinations of labels for the test sequence. (Even if $A(\zeta_1, \zeta_2)$ does not depend on the ordering of ζ_2, there are no computational savings unless the test sequence contains many identical objects.) We next introduce a family of region predictors based on the idea of the Bonferroni adjustment of p-values. In

brief, a Bonferroni predictor computes a p-value for each test object separately and then combines the k p-values into one p-value using the Bonferroni formula

$$p := \min(kp_1, \dots, kp_k, 1)\,. \tag{11}$$

The full description of the *Bonferroni predictor* (*BP*) corresponding to a nonconformity measure A is as follows:

- For each object x_j, $j \in \{l+1, \dots, l+k\}$, in the test sequence and each possible label $\upsilon \in \mathbf{Y}$:
  - set $y_j := \upsilon$ and $z_j := (x_j, y_j)$;
  - compute the nonconformity scores

$$\alpha_i := A((z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_l, z_j), z_i), \qquad i = 1, \dots, l, \tag{12}$$

$$\alpha_j := A((z_1, \dots, z_l), z_j)\,; \tag{13}$$

  - compute the p-value

$$p_{j-l}(\upsilon) := \frac{|\{i = 1, \dots, l \mid \alpha_i \geq \alpha_j\}| + 1}{l+1}\,. \tag{14}$$

- Output the prediction region

$$\Gamma^\epsilon(z_1, \dots, z_l, x_{l+1}, \dots, x_{l+k}) := \prod_{j=l+1}^{l+k} \{\upsilon \mid p_{j-l}(\upsilon) > \epsilon/k\}\,, \tag{15}$$

  where $\epsilon \in (0,1)$ is the significance level.

Notice that the prediction region (15) output by the BP can be rewritten in the form (2) if we define

$$p(\upsilon_1, \dots, \upsilon_k) := \min(kp_1(\upsilon_1), \dots, kp_k(\upsilon_k), 1) \tag{16}$$

(cf. (11)).

It is difficult to compare the rankmax TCP and the corresponding BP theoretically, but the following intermediate notion facilitates a comparison. The *semi-Bonferroni predictor* (*SBP*) is defined as follows:

- For each possible sequence of labels $(\upsilon_1, \dots, \upsilon_k) \in \mathbf{Y}^k$ for the test sequence:
  - set $y_j := \upsilon_{j-l}$ and $z_j := (x_j, y_j)$ for $j = l+1, \dots, l+k$;
  - compute nonconformity scores α_i, $i = 1, \dots, l+k$, by

$$\alpha_i := A((z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_{l+k}), z_i), \quad i = 1, \dots, l+k \tag{17}$$

    (cf. (7) and (4); the main difference of (17) from (7) and (4) is that (17) involves the true training observations and test objects whereas (7) and (4) involve arbitrary subsets of size l and k of the union of the training sequence and the test sequence with postulated labels);

– compute the p-value (14) for each $j = l+1, \dots, l+k$ and merge these p-values using (16).

- Output the prediction region (2).

Notice that the SBP becomes identical to the BP when (17) is replaced by (12) and (13) for $j = l+1, \dots, l+k$.

The following lemma shows that SBPs are usually weaker than the corresponding rankmax TCPs. (However, in Remark 3 and Section 6 we will see that the difference can be surprisingly small.)

Lemma 1. *Suppose all nonconformity scores (17) are different. The p-value (10) produced by a rankmax TCP does not exceed the p-value (16) produced by the corresponding SBP.*

Proof. Let t be the value of the rankmax test statistic, as defined at the end of Section 2. We are required to prove

$$1 - \frac{\binom{t-1}{k}}{\binom{l+k}{k}} \le k \frac{l+k-t+1}{l+1} . \tag{18}$$

Indeed, the left-hand side of (18) is identical to (10), and the ratio on the right-hand side of (18) is the smallest of the p-values (14) over j (cf. (16)). The statement that the ratio on the right-hand side of (18) is the smallest of the p-values (14) depends on (17) being all different (in fact, it is sufficient to assume that the maximum in the definition of the rankmax test statistic t is attained on only one test object). Notice, however, that the right-hand side of (18) is always an upper bound on the SBP p-value; this fact will be used in our discussions below.

We will prove a slightly stronger inequality than (18) replacing the denominator $l+1$ by $l+k$. In principle, t can take any value in $\{1, \dots, l+k\}$, but we can assume, without loss of generality, that $t \in \{k+1, \dots, l+k\}$: if $t \le k$, the left-hand side of (18) is 1 by definition and the right-hand side is at least 1 (even when $l+1$ is replaced by $l+k$). Rewriting (18) (with $l+k$ in place of $l+1$) as

$$1 - \frac{(t-1)(t-2)\cdots(t-k)}{k!\binom{l+k}{k}} \le k \frac{l+k-t+1}{l+k} , \tag{19}$$

we can assume that $t \in [k+1, l+k]$. Since the fraction on the left-hand side of (19) is a convex function of t (the second derivative is obviously nonnegative) and for $t := k+l$ (19) holds (it becomes $k/(l+k) \le k/(l+k)$), it suffices to prove that the derivative in t of the left-hand side of (19) at the point $l+k$ is equal to or exceeds the derivative of the right-hand side:

$$-\frac{(\Gamma(t)/\Gamma(t-k))'_{t=l+k}}{k!\binom{l+k}{k}} \ge k \frac{-1}{l+k} ,$$

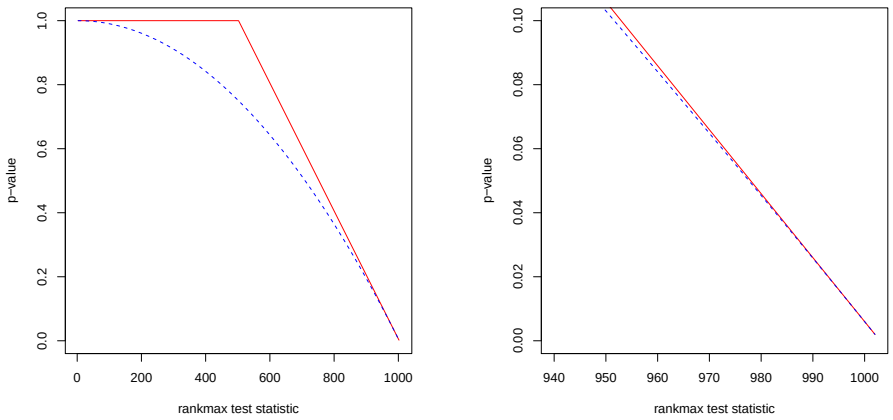

Figure 1: Left panel: SBP p-values (the solid red line) and rankmax TCP p-values (the dashed blue line) for $l = 1000$ and $k = 2$ as functions of t. Right panel: the lower right corner of the left panel.

where Γ is the gamma function, $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. By the definition of the digamma function ψ, the last inequality can be rewritten as

$$\frac{\Gamma(l+k)}{\Gamma(l)}\big(\psi(l+k) - \psi(l)\big) \le \frac{k}{l+k} k! \binom{l+k}{k},$$

which simplifies to

$$\psi(l+k) - \psi(l) \le \frac{k}{l}.$$

The well-known expression for ψ at the integer values of its argument (see, e.g., Olver et al.[6], `http://dlmf.nist.gov/5.4.14`) allows us to rewrite the last inequality as

$$\frac{1}{l} + \frac{1}{l+1} + \cdots + \frac{1}{l+k-1} \le \frac{k}{l},$$

which is obviously true. □

Remark. The proof of Lemma 1 shows that the inequality (18) is strict whenever $k > 1$ (for $k = 1$ the two p-values coincide). Three factors contribute to its being strict: the SBP p-value is larger than the rankmax TCP p-value at $t = l + k$; as function of t, the SBP p-value has a steeper (negative) slope at $t = l + k$; besides, to the left of $t = l + k$ the SBP p-value goes in a straight line whereas the rankmax TCP p-value veers down. This is illustrated in Figure 1 for $l = 1000$ and $k = 2$ (typical values for our experiments reported in Section 6); the first two factors, however, are not noticeable.

It is plausible that a BP usually produces somewhat smaller p-values (and, therefore, somewhat smaller prediction regions) than the corresponding SBP: the only difference is that, when computing p-values, the SBP uses more test objects with arbitrarily assigned labels, and this may lead to a greater distortion of the nonconformity scores.

4 Validity

The strongest notion of validity for conformal and related predictors can be stated in the online mode. Suppose we are given a sequence of positive integer numbers $k_1, k_2, \ldots$ and the incoming sequence of observations is $z_1 = (x_1, y_1), z_2 = (x_2, y_2), \ldots$; set $l_n := \sum_{i=1}^{n} k_i$ (including $l_0 := 0$). At trial $n = 1, 2, \ldots$ of the online prediction protocol, Predictor predicts the k_n labels $y_{l_{n-1}+1}, \ldots, y_{l_n}$ given the l_{n-1} observations $z_1, \ldots, z_{l_{n-1}}$ and k_n objects $x_{l_{n-1}+1}, \ldots, x_{l_n}$. The prediction is a subset Γ_n of $\mathbf{Y}^{k_n}$. It can be *multiple* ($|\Gamma_n| > 1$), *singleton* ($|\Gamma_n| = 1$), or *empty* ($|\Gamma_n| = 0$). Predictor *makes an error* if $(y_{l_{n-1}+1}, \ldots, y_{l_n}) \notin \Gamma_n$.

In this and next sections we assume either that the sequence of observations $z_1, z_2, \ldots$ is infinite and the observations are produced independently from the same probability distribution on $\mathbf{Z}$, or that the sequence of observations is finite, $z_1, \ldots, z_{l_N}$, and produced from an exchangeable probability distribution on $\mathbf{Z}^{l_N}$.

The following simple result states the validity of TCPs in the online mode; the idea of its proof is standard (see, e.g., Vovk[11] or Vovk et al.[13], Section 8.7).

Theorem 1. *In the online mode, a smoothed TCP makes errors with probability ϵ (the significance level) independently at different trials.*

In other words, Theorem 1 says that the sequence $e_1, e_2, \ldots$ of errors (where $e_i \in \{0, 1\}$ and $e_i = 1$ means that an error is made at trial i) is distributed as B_ϵ^∞ (or B_ϵ^N, if there are only l_N observations), where B_ϵ is the Bernoulli distribution with parameter (probability of success) equal to ϵ.

Proof. We will follow the scheme of the proof in Appendix A.1 of Vovk[11]. First notice that it is sufficient to prove that the sequence of errors $e_1, \ldots, e_n$ is distributed as B_ϵ^n for each n (for each $n \le N$ if there are l_N observations). Fix such an n. Given the multiset $\wr z_1, \ldots, z_{l_n} \int$ and under the assumption of exchangeability, the probability that the smoothed TCP will make an error at trial n is ϵ (this follows from the fact that p-values are distributed uniformly in the interval $[0, 1]$). Therefore, e_n is a Bernoulli random variable with parameter ϵ, even given the multiset $\wr z_1, \ldots, z_{l_n} \int$. Analogously, e_{n-1} is a Bernoulli random variable with parameter ϵ given the multiset $\wr z_1, \ldots, z_{l_n} \int$ and the sequence $(z_{l_{n-1}+1}, \ldots, z_{l_n})$; this implies that e_{n-1} is a Bernoulli random variable with parameter ϵ given e_n (since whether an error is made at trial n is determined by the multiset $\wr z_1, \ldots, z_{l_n} \int$ and observations $z_{l_{n-1}+1}, \ldots, z_{l_n}$: cf. Lemma 2 in Vovk[11]). Continuing backwards in this fashion, we eventually obtain that e_1

is a Bernoulli random variable with parameter ϵ given $e_2, \ldots, e_n$. Therefore, $e_1, \ldots, e_n$ is indeed distributed as B_ϵ^n. □

A suitable version of validity in the absence of smoothing is *conservative validity*, i.e., being dominated by a sequence of independent Bernoulli trials with parameter equal to the significance level. Formally (cf. Vovk et al.[13], p. 21), the conservative validity means that there is a probability space with two families

$$(\xi_n^{(\epsilon)} \mid \epsilon \in (0,1), n = 1, 2, \ldots), \;\; (\eta_n^{(\epsilon)} \mid \epsilon \in (0,1), n = 1, 2, \ldots)$$

of $\{0,1\}$-valued random variables such that:

- for a fixed ϵ, $\xi_1^{(\epsilon)}, \xi_2^{(\epsilon)}, \ldots$ is a sequence of independent Bernoulli random variables with parameter ϵ;
- for all n and ϵ, $\eta_n^{(\epsilon)} \leq \xi_n^{(\epsilon)}$;
- the joint distribution of errors $e_1, e_2, \ldots$ made at any significance level ϵ coincides with the joint distribution of $\eta_1^{(\epsilon)}, \eta_2^{(\epsilon)}, \ldots$.

By Theorem 1, TCPs are conservatively valid:

Corollary 1. *In the online mode, each TCP is conservatively valid.*

Proof. Each TCP is conservatively valid since it can only make an error when the corresponding smoothed TCP (i.e., the smoothed TCP based on the same transductive nonconformity measure) makes an error. □

Remark. Lemma 1 suggests that SBPs can be regarded as conservatively valid for practical purposes, since an SBP can make an error only when the corresponding rankmax TCP makes an error, unless there are ties among nonconformity scores. However, in general, it is not always true that an SBP can make an error only when the corresponding rankmax TCP makes an error. Consider, e.g., the case where $k = 2$ and the nonconformity scores of the two test observations are equal and exceed the nonconformity scores of all training observations; in this case, the SBP p-value will be smaller than the rankmax TCP p-value unless $l = 1$. Indeed, the required inequality between the rankmax TCP's p-value and the SBP's p-value is

$$1 - \frac{(l-1)l}{(l+1)(l+2)} \leq \frac{2}{l+1},$$

which is equivalent to $l \leq 1$.

Theorem 2. *In the online mode, each BP is conservatively valid.*

Proof. The proof follows the scheme of the proof of Theorem 1 above. Start, as before, by fixing an n. Given the multiset $\{\!|z_1, \ldots, z_{l_n}|\!\}$ and under the assumption of exchangeability, the probability that the BP will make an error at trial n for

a given test observation (e.g., for the second observation in the test sequence) is at most ϵ/k_n. Therefore, the probability that it will make an error for some of the k_n test observations is at most ϵ. We can increase the indicator e_n of making an error to obtain a Bernoulli random variable $\xi_n^{(\epsilon)}$ with parameter ϵ (this might involve extending the probability space). In this way we can obtain a sequence of independent $\xi_i^{(\epsilon)}$, $i = n, \dots, 1$, and it is clear that we can choose the same $\xi_i^{(\epsilon)}$ regardless of the starting value $n \geq i$. □

5 Universality

A *transductive confidence predictor* is a measurable strategy for Predictor in the online prediction protocol (as described in the previous section) depending on a parameter $\epsilon \in (0, 1)$ (the significance level) in such a way that for each training sequence and each test sequence the prediction at a larger significance level is a subset of the prediction at a smaller significance level. We say that the transductive confidence predictor is *conservatively valid* if the sequence of errors that it makes at any significance level ϵ is dominated by a sequence of independent Bernoulli trials with parameter ϵ. We say that it is *invariant* if, when fed with observations $z_1, \dots, z_{l_{n-1}}$ and objects $x_{l_{n-1}+1}, \dots, x_{l_n}$ at any trial n, it issues the same prediction regardless of the ordering of $z_1, \dots, z_{l_{n-1}}$. And we say that a transductive confidence predictor Γ' is *at least as good as* another transductive confidence predictor Γ'' if at any significance level ϵ the prediction region issued by Γ' is completely covered by the prediction region issued by Γ''. The following result states the universality of transductive conformal predictors.

Theorem 3. *Suppose* $\mathbf{Z}$ *is a Borel space. For any invariant conservatively valid transductive confidence predictor* Γ *there exists a transductive conformal predictor* Γ' *that is at least as good as* Γ.

Proof. The proof is similar to the proof of the analogous result for conformal predictors (Vovk et al.[13], Theorem 2.6). Fix an n ($n \leq N$ if there are l_N observations altogether). For each sequence $z_1, \dots, z_{l_n}$ of l_n observations define $f(z_1, \dots, z_{l_n})$ to be the conditional expectation of $\xi_n^{(\epsilon)}$ given $\eta_i^{(\epsilon)} = e_i$ for $i = 1, \dots, n$, where e_i is the indicator of error at trial i when the predictor Γ is fed with $z_1, \dots, z_{l_{i-1}}$. For each multiset B of size l_n let $f(B)$ be the arithmetic mean of f over all orderings of B. By the completeness of the statistic that maps a sequence of observations of length l_n to the corresponding multiset (Lehmann[3], Section 4.3), $f(B) = \epsilon$ for almost all multisets B; we will consider only such multisets. Define $S(B, \epsilon)$ as the multiset of sequences s in $\mathbf{Z}^{k_n}$ such that Γ makes an error at the significance level ϵ when fed with B ordered in such a way that it ends with s (because of the invariance of Γ, the order of the observations preceding s does not matter). Since

$$\epsilon_1 \leq \epsilon_2 \implies S(B, \epsilon_1) \subseteq S(B, \epsilon_2)$$

and

$$|S(B, \epsilon)| / l_n \leq \epsilon,$$

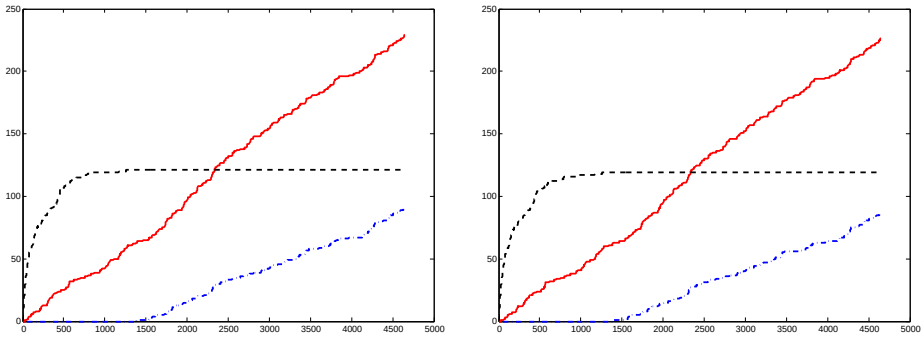

Figure 2: Left panel: the performance of the rankmax TCP based on Nearest Neighbour for tangent distance on the USPS data set (randomly permuted) for the length $k = 2$ of test sequences and significance level 5%. The cumulative errors are shown with a solid red line, multiple predictions with a dashed black line, and empty predictions with a dash-dot blue line. Right panel: the analogous picture for the BP.

the TCP Γ' corresponding to the transductive nonconformity measure

$$A((z_1, \dots, z_{l_{n-1}}), (z_{l_{n-1}+1}, \dots, z_{l_n})) := 1/\inf\left\{\epsilon \mid (z_{l_{n-1}+1}, \dots, z_{l_n}) \in S(\lfloor z_1, \dots, z_{l_n} \rfloor, \epsilon)\right\}$$

will be at least as good as Γ. □

Theorem 3 says that TCPs are universal in the sense of dominating all invariant conservatively valid transductive confidence predictors. In particular, for any BP there is a TCP that is at least as good as that BP. However, in the next section we will see that the rankmax TCP corresponding to the same nonconformity measure does not always satisfy this property.

6 Experiments

In our experiments we will use the standard USPS data set of hand-written digits. The training sequence (7291 observations) is merged with the test sequence (2007 observations) and the resulting sequence of 9298 observations is randomly permuted, to make sure the assumption of exchangeability is satisfied. The prediction protocol is online. In a typical scenario the digits might arrive in batches of $k = 5$ digits and represent American zip codes (in this case, however, the exchangeability assumption is only a crude approximation, since the digits within the same zip code are likely to be written by the same person). However, the TCP and SBP are too computationally inefficient to be applied in this case, and for comparing them with the BP we first consider online prediction of batches of $k = 2$ digits; intuitively, our task is to recognize a two-digit number.

We always use the Nearest Neighbour nonconformity measure (5), where d is tangent distance[8], and study empirically the corresponding rankmax TCP,

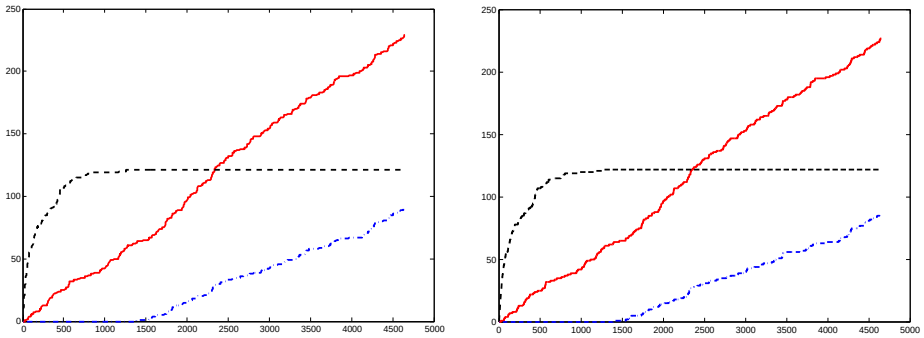

Figure 3: Left panel: reproduces the left panel of Figure 2 (the performance of the rankmax TCP). Right panel: the analogous picture for the SBP.

SBP, and BP. As the significance level we always take 5%. The left panel of Figure 2 shows the performance of the rankmax TCP using three functions: the cumulative number of errors made over the trials $1, \dots, n$ as function of n, the cumulative number of multiple predictions made over the trials $1, \dots, n$ as function of n, and the cumulative number of empty predictions over the trials $1, \dots, n$ as function of n. The performance of the SBP and BP as measured by these functions is very similar; only the latter is shown in the right panel of Figure 2, but all three graphs are visually indistinguishable (cf. Figure 3). The BP even makes 2 fewer multiple predictions than the rankmax TCP, which confirms the claim made in Section 5 that the rankmax TCP corresponding to the same nonconformity measure as a given BP is not always at least as good as that BP. (It is not true in general that the BP always makes fewer multiple predictions than the corresponding rankmax TCP. It just happens to be true for tangent distance and seed 0 for the MATLAB pseudorandom number generator; e.g., the BP makes slightly more multiple predictions for Euclidean distance and seed 0.) The SBP makes one more multiple prediction than the rankmax TCP, which agrees with Lemma 1.

The cause of the similarity between the two plots in Figure 3 is illustrated by Figure 4 (essentially a version of Figure 1), which shows the p-values produced by the SBP plotted against the respective p-values produced by the corresponding rankmax TCP, assuming there are no ties among the nonconformity scores. When the p-values are small, they are remarkably close to each other. And even without making any assumptions, we can still see that the SBP p-values are never significantly worse than the respective rankmax TCP p-values, assuming the latter are not too large.

The main advantage of BPs is that they are much more computationally efficient than both TCPs and SBPs. (For example, running the experiments described so far takes about 1 hour for the BP and about 50 hours for each of the TCP and the SBP on the computer system of a typical British department of computer science.) Because of their computational efficiency, it is very easy

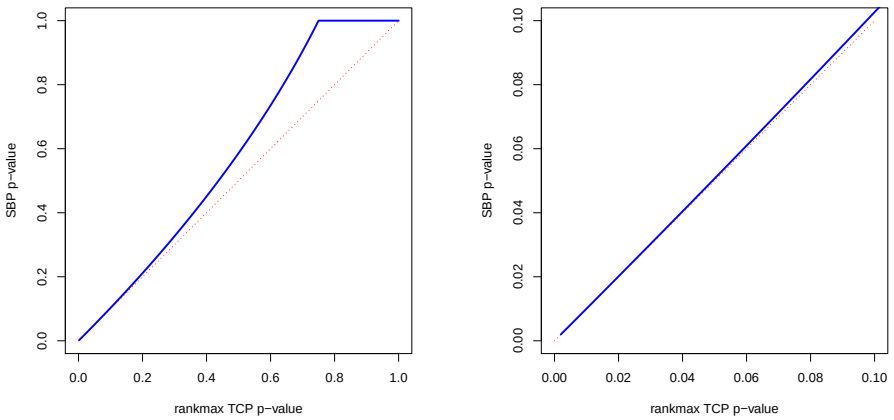

Figure 4: Left panel: the p-values produced by an SBP vs the p-values produced by the corresponding rankmax TCP (the solid blue line) for the length $l = 1000$ of the training sequence and $k = 2$ of the test sequence. Right panel: the lower left corner of the left panel.

to produce the analogue of the right panel of Figure 2 for $k = 5$ (as in American zip codes): see the left panel of Figure 5; but it is not clear at all how to make the computations for rankmax TCPs and SBPs feasible, even for moderately large k.

Remark. In our experiments in this section we only use transductive nonconformity measures $A : \mathbf{Z}^* \times \mathbf{Z}^* \to \mathbb{R}$ such that $A(\zeta_1, \zeta_2)$ depends on the ordering of neither ζ_1 nor ζ_2. In this case the definition of the p-value (1) can be rewritten, in obvious notation, as

$$p(\upsilon_1, \dots, \upsilon_k) := \frac{\#\{S \mid \alpha_S \geq \alpha_{\{l+1,\dots,l+k\}}\}}{\binom{l+k}{k}},$$

S ranging over all $\binom{l+k}{k}$ subsets of $\{1, \dots, l+k\}$ of size k.

Remark. It can be argued that in the main example of this section, recognizing zip codes, the cost of error depends on the position of the digit in the code: e.g., an error in the first digit will take the mail to a wrong state. This suggests the following generalization of BPs: the Bonferroni formula (11) is replaced by

$$p := \min(n_1 p_1, \dots, n_k p_k, 1),$$

where $n_1, \dots, n_k$ are positive constants satisfying

$$\frac{1}{n_1} + \dots + \frac{1}{n_k} = 1.$$

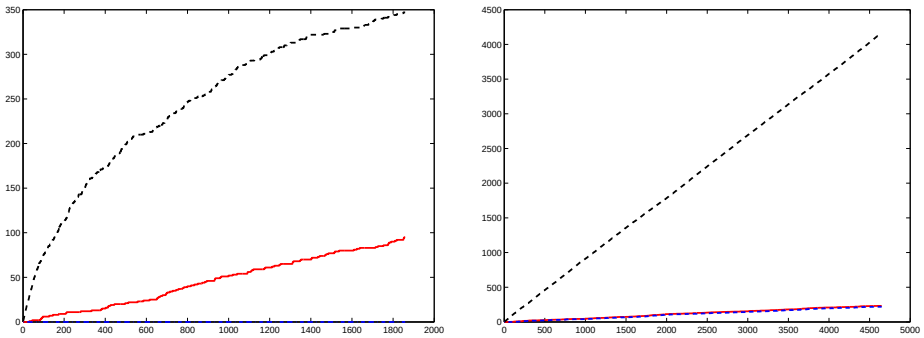

Figure 5: Left panel: the performance of the BP for the length $k = 5$ of test sequences. Right panel: the performance of the ranksum TCP for $k = 2$ (very poor). The setting is as in Figure 2: the prediction algorithms are based on Nearest Neighbour and tangent distance; the cumulative errors are shown with a solid red line, multiple predictions with a dashed black line, and empty predictions with a dash-dot blue line; the significance level is 5%.

(In other words, ϵ/k in (15) is replaced by ϵ/n_{j-l}.) For example, setting $n_1 = n_2 = n_3 = 10$, $n_4 = 5$, and $n_5 = 2$ makes the error in one of the digits (first, second, or third) determining the state unlikely; at significance level 5%, the probability of error in each of those digits is at most 0.5%.

7 Comparisons

The notion of transduction is usually opposed to induction. In this paper the relation between these two notions is more complicated, and they will be even combined in A into "transinductive conformal predictors". The goal of this section is to clarify our terminology.

Induction is the process of using training observations to arrive at a general rule; the general rule can then be applied to test objects to obtain predictions for their labels (deduction). Two essential properties of this process can be called informational and computational: the informational property is that we do not have the test set at the stage of finding the general rule, and the computational property is that the general rule is computationally efficient, so that applying it to new test objects can be done quickly (in many cases it can also be presented in a compact form). If a process of calculating predictions lacks one (or both) of these properties, it can be classified as transduction. Therefore, we have informational and computational transduction.

Vapnik and Chervonenkis[10] and Vapnik[9] emphasized informational transduction, where we should be given a test set before we can start computing predictions. Conformal predictors as defined in Vovk et al.[13] are not transductive in this sense. They are, however, transductive in the computational sense: for

each test object they have to redo most of the calculations. Their inductive version, inductive conformal predictors, was designed to lower computational burden.

In this paper, the word "induction" and its derivatives are always used in the computational sense, and the word "transduction" and its derivatives are always used in the informational sense.

For a further discussion of various aspects of transduction, see Vovk et al.[13], pp. 258–260.

8 Conclusion

Based on our theoretical and empirical results, the preliminary recommendation is to use Bonferroni predictors in transductive problems: as compared to rankmax TCPs and SBPs, they enjoy the same theoretical validity guarantees, have comparable predictive performance empirically, but are much more computationally efficient.

The conclusion is preliminary since our empirical comparison in Section 6 only covers TCPs for a small length k of the test sequence, namely $k = 2$. The computational inefficiency of TCPs greatly complicates their empirical comparison with the BPs and SBPs for large values of k.

The comparison is much more straightforward in the case of transductive and Bonferroni extensions of inductive conformal predictors (Papadopoulos et al.[7]; Vovk et al.[13], Section 4.1), and it can be shown that the two extensions produce similar p-values in practically important cases: see A for details.

The criterion of efficiency used in this paper for comparing various predictors in the transductive setting is the number of multiple predictions. This is a standard criterion (see, e.g., Vovk et al.[13]), but several recent papers[1, 2, 12] propose other criteria, which have certain advantages over the standard criterion. Comparison using the new criteria is a direction of further research.

Acknowledgments

I am grateful to Harris Papadopoulos for a discussion at COPA 2012 that rekindled my interest in transduction. Thanks to Wouter Koolen for illuminating discussions and for writing a MATLAB program for the Wilcoxon ranksum test (the standard programs in MATLAB and R produce unsatisfactory results in the default mode and are prohibitively slow in the exact mode). My thanks also go to the reviewers: the comments by the COPA 2013 reviewers helped me in improving the presentation and suggested new directions of research; the comments by the reviewers of the journal version of the paper have led to numerous further improvements, including the addition of Remark 6. In my experiments I used the C program for computing tangent distance written by Daniel Keysers and adapted to MATLAB by Aditi Krishn. This work was partially supported by the Cyprus Research Promotion Foundation (research

contract TPE/ORIZO/0609(BIE)/24), by EPSRC (grant EP/K033344/1), and by AFOSR (grant "Semantic completions").

References

[1] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Conformal prediction under hypergraphical models, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 9, 2013.

[2] Ulf Johansson, Rikard Konig, Tuve Lofstrom, and Henrik Bostrom. Evolved decision trees as conformal predictors. In Luis Gerardo de la Fraga, editor, *Proceedings of the 2013 IEEE Conference on Evolutionary Computation*, volume 1, pages 1794–1801, Cancun, Mexico, 2013.

[3] Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer, New York, second edition, 1986.

[4] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York, revised first edition, 2006.

[5] Ilia Nouretdinov, Sergi G. Costafreda, Alex Gammerman, Alexey Chervonenkis, Vladimir Vovk, Vladimir Vapnik, and Cynthia H. Y. Fu. Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *Neuroimage*, 56:809–813, 2011.

[6] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, 2010.

[7] 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*, pages 159–163, Las Vegas, NV, 2002. CSREA Press.

[8] 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.

[9] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[10] Vladimir N. Vapnik and Alexey Y. Chervonenkis. Теория распознавания образов *(Theory of Pattern Recognition)*. Nauka, Moscow, 1974. German translation: *Theorie der Zeichenerkennung*, Akademie, Berlin, 1979.

[11] Vladimir Vovk. On-line Confidence Machines are well-calibrated. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187–196, Los Alamitos, CA, 2002. IEEE Computer Society.

[12] Vladimir Vovk, Valentina Fedorova, Alex Gammerman, and Ilia Nouretdinov. Criteria of efficiency for conformal prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 11, April 2014.

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

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

[15] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

A Transinductive conformal predictors

Even BPs are computationally inefficient when used for predicting a large number of test sequences (of length k) from the same training sequence (of length l), since in general there is no way of reusing the computations carried out for the previous test sequences when processing the current test sequence. This appendix combines the ideas of TCPs and inductive conformal predictors (Papadopoulos et al.[7]; Vovk et al.[13], Section 4.1) to obtain a computationally efficient version of TCPs, which we will call transductive inductive, or transinductive, conformal predictors.

Split the training sequence $(z_1, \dots, z_l)$ into two parts: the *proper training sequence* $(z_1, \dots, z_m)$ of length $0 < m < l$ and the *calibration sequence* $(z_{m+1}, \dots, z_l)$ of length $l-m$; the test sequence $(x_{l+1}, \dots, x_{l+k})$ is as before. The *rankmax transinductive conformal predictor* (or *rankmax TICP*) corresponding to a nonconformity measure A is defined as follows:

- Compute the nonconformity scores

$$\alpha_i := A((z_1, \dots, z_m), z_i), \qquad i = m+1, \dots, l, \tag{20}$$

  for all calibration observations.

- For each possible sequence of labels $(\upsilon_1, \dots, \upsilon_k) \in \mathbf{Y}^k$:
  - set $y_j := \upsilon_{j-l}$ and $z_j := (x_j, y_j)$ for $j = l+1, \dots, l+k$;
  - compute the nonconformity scores

$$\alpha_j := A((z_1, \dots, z_m), z_j)\,, \tag{21}$$

    $j = l+1, \dots, l+k$, for all test observations;

– compute the p-value

$$p(\upsilon_1, \dots, \upsilon_k) := \frac{|\{S \mid \max_{i \in S} \alpha_i \geq \max(\alpha_{l+1}, \dots, \alpha_{l+k})\}|}{\binom{l-m+k}{k}}, \quad (22)$$

where S ranges over all $\binom{l-m+k}{k}$ subsets of $\{m+1, \dots, l+k\}$ of size k.

- Output the prediction region (2).

The *Bonferroni inductive predictor* (or *BIP*) corresponding to a nonconformity measure A is defined similarly:

- Compute the nonconformity scores (20) for all calibration observations.
- For each object x_j, $j \in \{l+1, \dots, l+k\}$, in the test sequence and each possible label $\upsilon \in \mathbf{Y}$:
  - set $y_j := \upsilon$ and $z_j := (x_j, y_j)$ for $j = l+1, \dots, l+k$;
  - compute the nonconformity score (21) for z_j;
  - compute the p-value

$$p_{j-l}(\upsilon) := \frac{|\{i = m+1, \dots, l \mid \alpha_i \geq \alpha_j\}| + 1}{l-m+1}.$$

- Output the prediction region (15).

The rankmax TICP and BIP are especially computationally efficient for nonconformity measures of the form (3), since the prediction rule f can be precomputed.

Another representation of the BIP is (2), where the p-values $p(\upsilon_1, \dots, \upsilon_k)$ are defined by (16). Lemma 1 simplifies in the inductive case:

Lemma 2. *If the nonconformity scores (21) for the test observations are all different, the p-value (22) produced by a rankmax TICP never exceeds the p-value (16) produced by the corresponding BIP.*

Proof. It suffices to apply (18) with $l-m$ (the length of the calibration sequence) in place of l to the value of the rankmax statistic $t := \max(R_{l+1}, \dots, R_{l+k})$, where R_j is now the rank of α_j in the multiset $\wr\alpha_{m+1}, \dots, \alpha_{l+k}\wr$. □

Section 6 was devoted to an empirical study of the difference between rankmax TCPs and the corresponding BPs. In the inductive case, such a study is, to a large degree, redundant. In the proof of Lemma 2 we saw that

$$1 - \frac{\binom{t-1}{k}}{\binom{l-m+k}{k}} \leq k \frac{l-m+k-t+1}{l-m+1},$$

where the left-hand side is the p-value produced by the rankmax TICP and the right-hand side is an upper bound on the p-value produced by the BIP (now

we are not making any assumptions on the nonconformity scores). When all nonconformity scores are different, the left panel of Figure 4 is also the plot of BIP p-values vs rankmax TICP p-values for a calibration sequence of length $l - m = 1000$ and a test sequence of length $k = 2$; Figure 1 and the discussion in Remark 3 are also applicable in the inductive case. When no assumptions are made, the pair of a rankmax p-value and the corresponding BIP p-value always lies at or below the solid blue line in Figure 4. We can see that the BIP's results are never much worse in the interesting range of small p-values. Figure 6 gives analogous pictures for the lengths $k = 10$ and $k = 50$ of the test sequence, and we still observe the same phenomenon: the BIP's results are never much worse if we are interested in small p-values; but the figure also illustrates the increasing difficulty of obtaining small p-values as the length k of the test sequence increases: it is clear that the smallest achievable p-value is $k/(l - m + 1)$ for the BIP and $k/(l - m + k)$ for the rankmax TICP.

Remark. The methods of transductive prediction considered in this paper can also be applied to cross-conformal predictors[14]; however, we will not go into the details of the resulting "transductive cross-conformal predictors".

B Ranksum TCP

This appendix briefly discusses ranksum TCPs, based on the ranksum aggregator (9). The results are shown in the right panel of Figure 5, in the same format as before. They are very poor, and the following heuristic argument explains why.

Suppose the training sequence is very long, of length $l \gg 1$, and the test sequence contains two observations. The TCP assigns all possible labels to the two test objects, and we can expect the prediction to be a singleton whenever assigning a wrong label to either test object leads to a p-value not exceeding the significance level. Now suppose one of the test objects is assigned the correct label and the other a wrong label. Let us assume, optimistically, that the normalized rank $R/(l + 2)$ of the latter test object (with a wrong label) is 1; the normalized rank x of the former test object (with the right label) will be, at best, uniformly distributed on $[0, 1]$. In the limit of a very long training sequence and assuming the observations are exchangeable, the p-value corresponding to the normalized rank x of the former test object is at least

$$\mathbb{P}(\xi_1 + \xi_2 \geq 1 + x) = (1 - x)^2/2\,,$$

where ξ_1 and ξ_2 are distributed uniformly on $[0, 1]$, and so the expected p-value is at least $\int_0^1 \frac{(1-x)^2}{2} dx = 1/6 \approx 17\%$.

This shows that we can expect the bulk of our predictions to be singleton when using the ranksum TCP only when the significance level considerably exceeds 17%. For example, the ranksum TCP for the USPS data set at the 5% level will typically produce as its prediction the cross in $\{0, \dots, 9\}^2$ centred on the pair of true labels, and this has been observed in our experiments.

Remark. In fact, the results for the most straightforward implementation of the ranksum TCP are unsatisfactory for two reasons: the very low accuracy of computed p-values for the Wilcoxon ranksum test (both in MATLAB and R) and the test's way of combining ranks (summing) that is very poorly suited to our task. Our argument shows that already the second reason is sufficient, and this is confirmed by more precise calculations of p-values for the Wilcoxon ranksum test.

Remark. It is interesting that using an unsuitable rank aggregator leads to the predictor that sometimes issues empty predictions before multiple predictions (see the right panel of Figure 5). Typically satisfactory predictors start issuing empty predictions only after they stop issuing multiple predictions. For example, the rankmax TCP in the left panel of Figure 2 issues only singleton predictors in trials 1263–1385; before that it never issues empty predictions and after that it never issues multiple predictions.

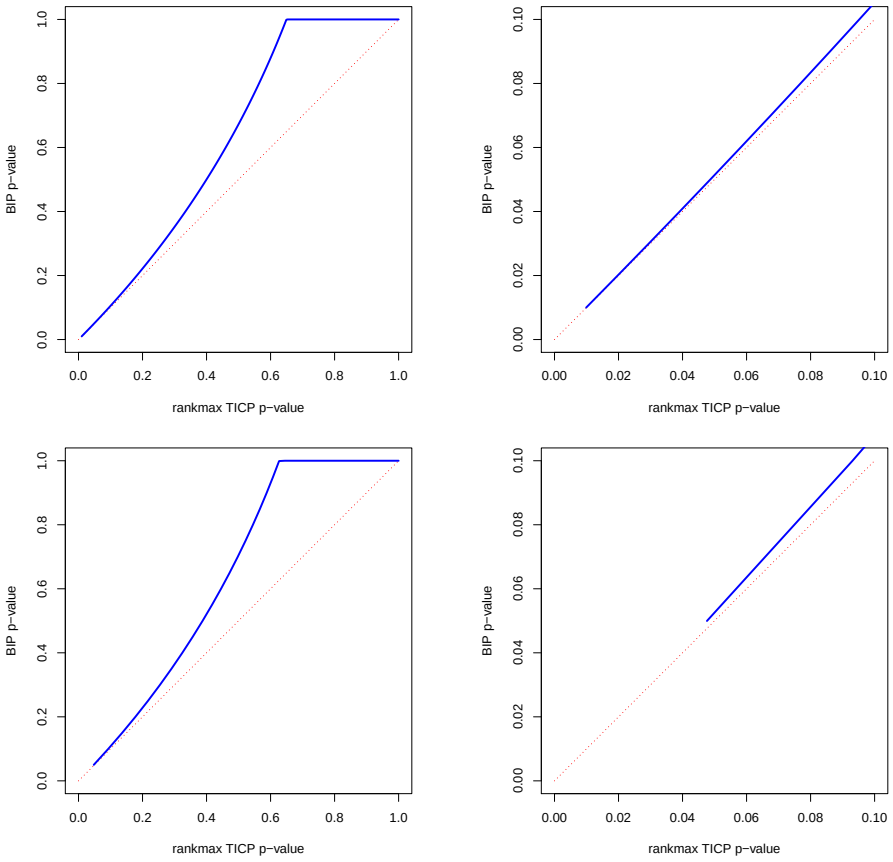

Figure 6: Left panels: the p-values produced by a BIP vs the p-values produced by the corresponding rankmax TICP (the solid blue lines). Right panels: the lower left corners of the corresponding left panels. The length of the calibration sequence is $l - m = 1000$ and the length of the test sequence is $k = 10$ for the top panels and $k = 50$ for the bottom panels.