From conformal to probabilistic prediction

Vladimir Vovk, Ivan Petej, and Valentina Fedorova

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 #12

First posted May 11, 2014. Last revised January 19, 2019.

Project web site:
http://alrw.net

Abstract

This paper proposes a new method of probabilistic prediction, which is based on conformal prediction. The method is applied to the standard USPS data set and gives encouraging results.

Contents

1 Introduction **1**

2 Criteria of efficiency for label-conditional conformal predictors and transducers **2**

3 Optimal idealized conformity measures for a known probability distribution **3**

4 Criteria of efficiency for probabilistic predictors **6**

5 Calibration of p-values into conditional probabilities **7**

6 Experiments **8**

References **10**

1 Introduction

In essence, conformal predictors output systems of p-values: to each potential label of a test object a conformal predictor assigns the corresponding p-value, and a low p-value is interpreted as the label being unlikely. It has been argued, especially by Bayesian statisticians, that p-values are more difficult to interpret than probabilities; besides, in decision problems probabilities can be easily combined with utilities to obtain decisions that are optimal from the point of view of Bayesian decision theory. In this paper we will apply the idea of transforming p-values into probabilities (used in a completely different context in, e.g., [10, Sect. 9] and [7]) to conformal prediction: the p-values produced by conformal predictors will be transformed into probabilities.

The approach of this paper is as follows. It was observed in [12] that some criteria of efficiency for conformal prediction (called "probabilistic criteria") encourage using the conditional probability $Q(y \mid x)$ as the conformity score for an observation (x, y), Q being the data-generating distribution. In this paper we extend this observation to label-conditional predictors (Sect. 2).

Next we imagine that we are given a conformal predictor Γ that is nearly optimal with respect to a probabilistic criterion (such a conformal predictor might be an outcome of a thorough empirical study of various conformal predictors using a probabilistic criterion of efficiency). Essentially, this means that in the limit of a very large training set the p-value that Γ outputs for an observation (x, y) is a monotonic transformation of the conditional probability $Q(y \mid x)$ (Theorem 1 in Sect. 3).

Finally, we transform the p-values back into conditional probabilities using the distribution of p-values in the test set (Sect. 5). Following [10] and [7], we will say that at this step we *calibrate* the p-values into probabilities,

In Sect. 6 we give an example of a realistic situation where use of the techniques developed in this paper improves on a standard approach. The performance of the probabilistic predictors considered in that section is measured using standard loss functions, logarithmic and Brier (Sect. 4).

Comparisons with related work

It should be noted that in the process of transforming p-values into probabilities suggested in this paper we lose a valuable feature of conformal prediction, its automatic validity. Our hope, however, is that the advantages of conformal prediction will translate into accurate probabilistic predictions.

There is another method of probabilistic prediction that is related to conformal prediction, Venn prediction (see, e.g., [13, Chap. 6] or [14]). This method does have a guaranteed property of validity (perhaps the simplest being Theorem 1 in [14]); however, the price to pay is that it outputs multiprobabilistic predictions rather than sharp probabilistic predictions. There are natural ways of transforming multiprobabilistic predictions into sharp probabilistic predictions (see, e.g., [14, Sect. 4]), but such transformations, again, lead to the loss of the formal property of validity.

As preparation, we study label-conditional conformal prediction. For a general discussion of conditionality in conformal prediction, see [11]. Object-conditional conformal prediction has been studied in [5] (in the case of regression).

2 Criteria of efficiency for label-conditional conformal predictors and transducers

Let $\mathbf{X}$ be a measurable space (the *object space*) and $\mathbf{Y}$ be a finite set equipped with the discrete σ-algebra (the *label space*); the *observation space* is defined to be $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$. A *conformity measure* is a measurable function A that assigns to every sequence $(z_1, \dots, z_n) \in \mathbf{Z}^*$ of observations a same-length sequence $(\alpha_1, \dots, \alpha_n)$ of real numbers and that is equivariant with respect to permutations: for any n and any permutation π of $\{1, \dots, n\}$,

$$(\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n) \Longrightarrow \left(\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}\right) = A\left(z_{\pi(1)}, \dots, z_{\pi(n)}\right).$$

The *label-conditional conformal predictor* determined by A is defined by

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

where $(z_1, \dots, z_l) \in \mathbf{Z}^*$ is a training sequence, x is a test object, $\epsilon \in (0, 1)$ is a given *significance level*, and for each $y \in \mathbf{Y}$ the corresponding *label-conditional p-value* p^y is defined by

$$p^y := \frac{\left|\left\{i = 1, \dots, l+1 \mid y_i = y \ \ \& \ \ \alpha_i^y < \alpha_{l+1}^y\right\}\right|}{\left|\{i = 1, \dots, l+1 \mid y_i = y\}\right|} + \tau \frac{\left|\left\{i = 1, \dots, l+1 \mid y_i = y \ \ \& \ \ \alpha_i^y = \alpha_{l+1}^y\right\}\right|}{\left|\{i = 1, \dots, l+1 \mid y_i = y\}\right|}, \tag{2}$$

where τ is a random number distributed uniformly on the interval $[0, 1]$ and the corresponding sequence of *conformity scores* is defined by

$$(\alpha_1^y, \dots, \alpha_l^y, \alpha_{l+1}^y) := A(z_1, \dots, z_l, (x, y)).$$

It is clear that the system of *prediction sets* (1) output by a conformal predictor is nested, namely decreasing in ϵ.

The *label-conditional conformal transducer* determined by A outputs the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (2) for each training sequence $(z_1, \dots, z_l)$ of observations and each test object x.

Four criteria of efficiency

Suppose that, besides the training sequence, we are also given a test sequence, and would like to measure on it the performance of a label-conditional conformal predictor or transducer. As usual, let us define the performance on the test set

to be the average performance (or, equivalently, the sum of performances) on the individual test observations. Following [12], we will discuss the following four criteria of efficiency for individual test observations; all the criteria will work in the same direction: the smaller the better.

- The sum $\sum_{y\in\mathbf{Y}} p^y$ of the p-values; referred to as the *S criterion*. This is applicable to conformal transducers (i.e., the criterion is ϵ-independent).
- The size $|\Gamma^\epsilon|$ of the prediction set at a significance level ϵ; this is the *N criterion*. It is applicable to conformal predictors (ϵ-dependent).
- The sum of the p-values apart from that for the true label: the *OF* ("observed fuzziness") *criterion*.
- The number of false labels included in the prediction set Γ^ϵ at a significance level ϵ; this is the *OE* ("observed excess") *criterion*.

The last two criteria are simple modifications of the first two (leading to smoother and more expressive pictures).

Remark 1. Equivalently, the S criterion can be defined as the arithmetic mean $\frac{1}{|\mathbf{Y}|}\sum_{y\in\mathbf{Y}} p^y$ of the p-values; the proof of Theorem 1 below will show that, in fact, we can replace arithmetic mean by any mean [3, Sect. 3.1], including geometric, harmonic, etc.

3 Optimal idealized conformity measures for a known probability distribution

In this section we consider the idealized case where the probability distribution Q generating independent observations $z_1, z_2, \ldots$ is known (as in [12]). The main result of this section, Theorem 1, is the label-conditional counterpart of Theorem 1 in [12]; the proof of our Theorem 1 is also modelled on the proof of Theorem 1 in [12]. In this section we assume, for simplicity, that the set $\mathbf{Z}$ is finite and that $Q(\{z\}) > 0$ for all $z \in \mathbf{Z}$.

An *idealized conformity measure* is a function $A(z, Q)$ of $z \in \mathbf{Z}$ and $Q \in \mathcal{P}(\mathbf{Z})$ (where $\mathcal{P}(\mathbf{Z})$ is the set of all probability measures on $\mathbf{Z}$). We will sometimes write the corresponding conformity scores as $A(z)$, as Q will be clear from the context. The *idealized smoothed label-conditional conformal predictor* corresponding to A outputs the following prediction set $\Gamma^\epsilon(x)$ for each object $x \in \mathbf{X}$ and each significance level $\epsilon \in (0, 1)$. For each potential label $y \in \mathbf{Y}$ for x define the corresponding *label-conditional p-value* as

$$p^y = p(x, y) := \frac{Q(\{(x', y) \mid x' \in \mathbf{X} \ \&\ A((x', y), Q) < A((x, y), Q)\})}{Q_{\mathbf{Y}}(\{y\})} + \tau \frac{Q(\{(x', y) \mid x' \in \mathbf{X} \ \&\ A((x', y), Q) = A((x, y), Q)\})}{Q_{\mathbf{Y}}(\{y\})} \quad (3)$$

(this is the idealized analogue of (2)), where $Q_{\mathbf{Y}}$ is the marginal distribution of Q on $\mathbf{Y}$ and τ is a random number distributed uniformly on $[0,1]$. The prediction set is

$$\Gamma^\epsilon(x) := \{y \in \mathbf{Y} \mid p(x,y) > \epsilon\}. \tag{4}$$

The *idealized smoothed label-conditional conformal transducer* corresponding to A outputs for each object $x \in \mathbf{X}$ the system of p-values $(p^y \mid y \in \mathbf{Y})$ defined by (3); in the idealized case we will usually use the alternative notation $p(x,y)$ for p^y.

Four idealized criteria of efficiency

In this subsection we will apply the four criteria of efficiency that we discussed in the previous section to the idealized case of infinite training and test sequences; since the sequences are infinite, they carry all information about the data-generating distribution Q. We will write $\Gamma^\epsilon_A(x)$ for the $\Gamma^\epsilon(x)$ in (4) and $p_A(x,y)$ for the $p(x,y)$ in (3) to indicate the dependence on the choice of the conformity measure A. Let U be the uniform probability measure on the interval $[0,1]$.

An idealized conformity measure A is:

- *S-optimal* if $\mathbb{E}_{(x,\tau)\sim Q_{\mathbf{X}}\times U} \sum_y p_A(x,y) \le \mathbb{E}_{(x,\tau)\sim Q_{\mathbf{X}}\times U} \sum_y p_B(x,y)$ for any idealized conformity measure B, where $Q_{\mathbf{X}}$ is the marginal distribution of Q on $\mathbf{X}$;
- *N-optimal* if $\mathbb{E}_{(x,\tau)\sim Q_{\mathbf{X}}\times U} |\Gamma^\epsilon_A(x)| \le \mathbb{E}_{(x,\tau)\sim Q_{\mathbf{X}}\times U} |\Gamma^\epsilon_B(x)|$ for any idealized conformity measure B and any significance level ϵ;
- *OF-optimal* if

  $$\mathbb{E}_{((x,y),\tau)\sim Q\times U} \sum_{y'\ne y} p_A(x,y') \le \mathbb{E}_{((x,y),\tau)\sim Q\times U} \sum_{y'\ne y} p_A(x,y')$$

  for any idealized conformity measure B;
- *OE-optimal* if

  $$\mathbb{E}_{((x,y),\tau)\sim Q\times U} |\Gamma^\epsilon_A(x) \setminus \{y\}| \le \mathbb{E}_{((x,y),\tau)\sim Q\times U} |\Gamma^\epsilon_B(x) \setminus \{y\}|$$

  for any idealized conformity measure B and any significance level ϵ.

The *conditional probability (CP) idealized conformity measure* is

$$A((x,y),Q) := Q(y \mid x).$$

An idealized conformity measure A is a (label-conditional) *refinement* of an idealized conformity measure B if

$$B((x_1,y)) < B((x_2,y)) \Longrightarrow A((x_1,y)) < A((x_2,y))$$

for all $x_1, x_2 \in \mathbf{Z}$ and all $y \in \mathbf{Y}$. (Notice that this definition, being label-conditional, is different from the one given in [12].) Let $\mathcal{R}(\mathrm{CP})$ be the set of all refinements of the CP idealized conformity measure. If C is a criterion of efficiency (one of the four discussed above), we let $\mathcal{O}(C)$ stand for the set of all C-optimal idealized conformity measures.

Theorem 1. $\mathcal{O}(\mathrm{S}) = \mathcal{O}(\mathrm{OF}) = \mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{OE}) = \mathcal{R}(\mathrm{CP})$.

Proof. We start from proving $\mathcal{R}(\mathrm{CP}) = \mathcal{O}(\mathrm{N})$. Fix a significance level ϵ. A smoothed confidence predictor at level ϵ is defined as a random set of observations $(x, y) \in \mathbf{Z}$; in other words, to each observation (x, y) is assigned the probability $P(x, y)$ that the observation will be outside the prediction set. Under the restriction that the sum of the probabilities $Q(x, y)$ of observations (x, y) outside the prediction set (defined as $\sum_x Q(x, y)P(x, y)$ in the smoothed case) is bounded by $\epsilon Q_{\mathbf{Y}}(y)$ for a fixed y, the N criterion requires us to make the sum of $Q_{\mathbf{X}}(x)$ for (x, y) outside the prediction set (defined as $\sum_x Q_{\mathbf{X}}(x)P(x, y)$ in the smoothed case) as large as possible. It is clear that the set should consist of the observations with the smallest $Q(y \mid x)$ (by the usual Neyman–Pearson argument: cf. [4, Sect. 3.2]).

Next we show that $\mathcal{O}(\mathrm{N}) \subseteq \mathcal{O}(\mathrm{S})$. Let an idealized conformity measure A be N-optimal. By definition,

$$\mathbb{E}_{x,\tau} \left|\Gamma^\epsilon_A(x)\right| \le \mathbb{E}_{x,\tau} \left|\Gamma^\epsilon_B(x)\right|$$

for any idealized conformity measure B and any significance level ϵ. Integrating over $\epsilon \in (0, 1)$ and swapping the order of integrals and expectations,

$$\mathbb{E}_{x,\tau} \int_0^1 \left|\Gamma^\epsilon_A(x)\right| \, \mathrm{d}\epsilon \le \mathbb{E}_{x,\tau} \int_0^1 \left|\Gamma^\epsilon_B(x)\right| \, \mathrm{d}\epsilon. \tag{5}$$

Since

$$\left|\Gamma^\epsilon(x)\right| = \sum_{y \in \mathbf{Y}} 1_{\{p(x,y) > \epsilon\}},$$

we can rewrite (5), after swapping the order of summation and integration, as

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} \left(\int_0^1 1_{\{p_A(x,y) > \epsilon\}} \, \mathrm{d}\epsilon \right) \le \mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} \left(\int_0^1 1_{\{p_B(x,y) > \epsilon\}} \, \mathrm{d}\epsilon \right).$$

Since

$$\int_0^1 1_{\{p(x,y) > \epsilon\}} \, \mathrm{d}\epsilon = p(x, y),$$

we finally obtain

$$\mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_A(x, y) \le \mathbb{E}_{x,\tau} \sum_{y \in \mathbf{Y}} p_B(x, y).$$

Since this holds for any idealized conformity measure B, A is S-optimal.

The argument in the previous paragraph in fact shows that $\mathcal{O}(\mathrm{S}) = \mathcal{O}(\mathrm{N}) = \mathcal{R}(\mathrm{CP})$. Indeed, that argument shows that

$$\sum_{y\in\mathbf{Y}} p(x,y) = \int_0^1 |\Gamma^\epsilon(x)| \,\mathrm{d}\epsilon,$$

and so to optimize a conformity measure in the sense of the S criterion it suffices to optimize it in the sense of the N criterion for all ϵ simultaneously (which can, and therefore should, be done). More generally, for any continuous increasing function ϕ we have

$$\begin{aligned}\sum_{y\in\mathbf{Y}} \phi(p(x,y)) = \sum_{y\in\mathbf{Y}} \int_0^1 1_{\{\phi(p(x,y))>\epsilon\}} \,\mathrm{d}\epsilon = \int_0^1 \sum_{y\in\mathbf{Y}} 1_{\{p(x,y)>\phi^{-1}(\epsilon)\}} \,\mathrm{d}\epsilon \\ = \int_0^1 \left|\Gamma^{\phi^{-1}(\epsilon)}(x)\right| \,\mathrm{d}\epsilon = \int \left|\Gamma^{\epsilon'}(x)\right| \phi'(\epsilon') \,\mathrm{d}\epsilon',\end{aligned}$$

which proves Remark 1.

The equality $\mathcal{O}(\mathrm{S}) = \mathcal{O}(\mathrm{OF})$ follows from

$$\mathbb{E}_{x,\tau} \sum_y p(x,y) = \mathbb{E}_{(x,y),\tau} \sum_{y'\neq y} p(x,y') + \frac{1}{2},$$

where we have used the fact that $p(x,y)$ is distributed uniformly on $[0,1]$ when $((x,y),\tau) \sim Q \times U$ (see [13] and [12]).

Finally, we notice that $\mathcal{O}(\mathrm{N}) = \mathcal{O}(\mathrm{OE})$. Indeed, for any significance level ϵ,

$$\mathbb{E}_{x,\tau} |\Gamma^\epsilon(x)| = \mathbb{E}_{(x,y),\tau} |\Gamma^\epsilon(x) \setminus \{y\}| + (1-\epsilon),$$

again using the fact that $p(x,y)$ is distributed uniformly on $[0,1]$ and so $\mathbb{P}_{(x,y),\tau}(y \in \Gamma^\epsilon(x)) = 1-\epsilon$. □

4 Criteria of efficiency for probabilistic predictors

Given a training set $(z_1,\dots,z_l)$ and a test object x, a probabilistic predictor outputs a probability measure $P \in \mathcal{P}(\mathbf{Y})$, which is interpreted as its probabilistic prediction for the label y of x; we let $\mathcal{P}(\mathbf{Y})$ stand for the set of all probability measures on $\mathbf{Y}$. The two standard way of measuring the performance of P on the actual label y are the *logarithmic* (or *log*) *loss* $-\ln P(\{y\})$ and the *Brier loss*

$$\sum_{y'\in\mathbf{Y}} \left(1_{\{y'=y\}} - P(\{y'\})\right)^2,$$

where 1_E stands for the indicator of an event E: $1_E = 0$ if E happens and $1_E = 0$ otherwise. The efficiency of probabilistic predictors will be measured by these two loss functions.

Suppose we have a test sequence $(z_{l+1}, \dots, z_{l+k})$, where $z_i = (x_i, y_i)$ for $i = l+1, \dots, l+k$, and we want to evaluate the performance of a probabilistic predictor (trained on a training sequence $z_1, \dots, z_l$) on it. In the next section we will use the *average log loss*

$$-\frac{1}{k} \sum_{i=l+1}^{l+k} \ln P_i(\{y_i\})$$

and the *standardized Brier loss*

$$\sqrt{\frac{1}{k\,|\mathbf{Y}|} \sum_{i=l+1}^{l+k} \sum_{y' \in \mathbf{Y}} \left(1_{\{y'=y_i\}} - P_i(\{y'\})\right)^2},$$

where $P_i \in \mathcal{P}(\mathbf{Y})$ is the probabilistic prediction for x_i. Notice that in the binary case, $|\mathbf{Y}| = 2$, the average log loss coincides with the mean log error (used in, e.g., [14, (12)]) and the standardized Brier loss coincides with the root mean square error (used in, e.g., [14, (13)]).

5 Calibration of p-values into conditional probabilities

The argument of this section will be somewhat heuristic, and we will not try to formalize it in this paper. Fix $y \in \mathbf{Y}$. Suppose that $q := P(y \mid x)$ has an absolutely continuous distribution with density f when $x \sim Q_{\mathbf{X}}$. (In other words, f is the density of the image of $Q_{\mathbf{X}}$ under the mapping $x \mapsto P(y \mid x)$.) For the CP idealized conformity measure, we can rewrite (3) as

$$p(q) := \int_0^q q' f(q') dq' \Big/ D \,, \tag{6}$$

where $D := Q_{\mathbf{Y}}(\{y\})$; alternatively, we can set $D := \int_0^1 q' f(q') dq'$ to the normalizing constant ensuring that $p(1) = 1$. To see how (6) is a special case of (3) for the CP idealized conformity measure, notice that the probability that $Y = y$ and $P(Y \mid X) \in (q', q' + dq')$, where $(X, Y) \sim f$, is $q' f(q') dq'$. In (6) we write $p(q)$ rather than p^y since p^y depends on y only via q.

We are more interested in the inverse function $q(p)$, which is defined by the condition

$$p = \int_0^{q(p)} q' f(q') dq' \Big/ D \,.$$

When $q \sim f$, we have

$$\mathbb{P}(p(q) \le a) = \mathbb{P}(q \le q(a)) = \int_0^{q(a)} f(q') dq'.$$

Algorithm 1 Conformal-type probabilistic predictor

Input: training sequence $(z_1, \dots, z_l) \in \mathbf{Z}^l$
Input: calibration sequence $(x_{l+1}, \dots, x_{l+k}) \in \mathbf{X}^k$
Input: test object x_0
Output: probabilistic prediction $P \in \mathcal{P}(\mathbf{Y})$ for the label of x_0
  for $y \in \mathbf{Y}$ **do**
    for each x_i in the calibration sequence find the p-value p_i^y by (2)
      (with $l+i$ in place of $l+1$)
    let g_y be the antitonic density on $[0,1]$ fitted to $p_{l+1}^y, \dots, p_{l+k}^y$
    find the p-value p_0^y by (2) (with 0 in place of $l+1$)
    for each $y \in \mathbf{Y}$, set $P'(\{y\}) := g_y(1)/g_y(p_0^y)$
  end for
  set $P(\{y\}) := P'(\{y\}) / \sum_{y'} P'(\{y'\})$ for each $y \in \mathbf{Y}$

Therefore, when $q \sim f$, we have

$$\mathbb{P}(a \le p(q) \le a + da) = \int_{q(a)}^{q(a+da)} f(q')dq' \approx \frac{1}{q(a)} \int_{q(a)}^{q(a+da)} q' f(q')dq' = \frac{Dda}{q(a)},$$

and so

$$q(c) \approx D \Big/ \frac{\mathbb{P}(c \le p(q) \le c + dc)}{dc}.$$

This gives rise to the algorithm given as Algorithm 1, which uses real p-values (2) instead of the ideal p-values (3). The algorithm is transductive in that it uses a training sequence of labelled observations and a calibration sequence of unlabelled objects (in the next section we use the test sequence as the calibration sequence); the latter is used for calibrating p-values into conditional probabilities. Given all the p-values for the calibration sequence with postulated label y, find the corresponding antitonic density $g(p)$ (remember that the function $q(p)$ is known to be monotonic, namely isotonic) using Grenander's estimator (see [2] or, e.g., [1, Chap. 8]). Use $D/g(p)$ as the calibration function, where $D := g(1)$ is chosen in such a way that a p-value of 1 is calibrated into a conditional probability of 1. (Alternatively, we could set D to the fraction of observations labelled as y in the training sequence; this approximates setting $D := Q_{\mathbf{Y}}(\{y\})$.) The probabilities produced by this procedure are not guaranteed to lead to a probability measure: the sum over y can be different from 1 (and this phenomenon has been observed in our experiments). Therefore, in the last line of Algorithm 1 we normalize the calibrated p-values to obtain genuine probabilities.

6 Experiments

In our experiments we use the standard USPS data set of hand-written digits. The size of the training set is 7291, and the size of the test set is 2007; however,

Table 1: The performance of the two algorithms, Platt's (with the optimal values of parameters) and the conformal-type probabilistic predictor based on 1-Nearest Neighbour with tangent distance

algorithm	average log loss	standardized Brier loss
optimized Platt	0.06431	0.05089
conformal-type 1-NN	0.04958	0.04359

instead of using the original split of the data into the two parts, we randomly split all available data (the union of the original training and test sets) into a training set of size 7291 and test set of size 2007. (Therefore, our results somewhat depend on the seed used by the random number generator, but the dependence is minor and does not affect our conclusions at all; we always report results for seed 0.)

A powerful algorithm for the USPS data set is the 1-Nearest Neighbour (1-NN) algorithm using tangent distance [8]. However, it is not obvious how this algorithm could be transformed into a probabilistic predictor. On the other hand, there is a very natural and standard way of extracting probabilities from support vector machines, which we will refer to it as *Platt's algorithm* in this paper: it is the combination of the method proposed by Platt [6] with pairwise coupling [15] (unlike our algorithm, which is applicable to multi-class problems directly, Platt's method is directly applicable only to binary problems). In this section we will apply our method to the 1-NN algorithm with tangent distance and compare the results to Platt's algorithm as implemented in the function `svm` from the `e1071` R package (for our multi-class problem this function calculates probabilities using the combination of Platt's binary method and pairwise coupling).

There is a standard way of turning a distance into a conformal predictor [13, Sect. 3.1]: namely, the conformity score α_i of the ith observation in a sequence of observations can be defined as

$$\frac{\min_{j:y_j \neq y_i} d(x_i, x_j)}{\min_{j \neq i: y_j = y_i} d(x_i, x_j)}, \tag{7}$$

where d is the distance; the intuition is that an object is considered conforming if it is close to an object labelled in the same way and far from any object labelled in a different way.

Table 1 compares the performance of the conformal-type probabilistic predictor based on the 1-NN conformity measure (7), where d is tangent distance, with the performance of Platt's algorithm with the optimal values of its parameters. The conformal predictor is parameter-free but Platt's algorithm depends on the choice of the kernel. We chose the polynomial kernel of degree 3 (since it is known to produce the best results: see [9, Sect. 12.2]) and the cost parameter $C := 2.9$ in the case of the average log loss and $C := 3.4$ in the case of the

Table 2: The performance of Platt's algorithm with the polynomial kernels of various degrees for the cost parameter $C = 10$

degree	average log loss	standardized Brier loss
1	0.12681	0.07342
2	0.09967	0.06109
3	0.06855	0.05237
4	0.11041	0.06227
5	0.09794	0.06040

standardized Brier loss (the optimal values in our experiments). (Reporting the performance of Platt's algorithm with optimal parameter values may look like data snooping, but it is fine in this context since we are helping our competitor.) Table 2 reports the performance of Platt's algorithm as function of the degree of the polynomial kernel with the cost parameter set at $C := 10$ (the dependence on C is relatively mild, and $C = 10$ gives good performance for all degrees that we consider).

Acknowledgments.

We thank the reviewer for useful comments. In our experiments we used the R package `e1071` (by David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, and Chih-Chen Lin) and the implementation of tangent distance by Daniel Keysers. This work was partially supported by EPSRC (grant EP/K033344/1, first author) and Royal Holloway, University of London (third author).

References

[1] Luc Devroye. *A Course in Density Estimation*. Birkhäuser, New York, 1987.

[2] Ulf Grenander. On the theory of mortality measurement. Part II. *Skandinavisk Aktuarietidskrift*, 39:125–153, 1956.

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

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

[5] Jing Lei and Larry Wasserman. Distribution free prediction bands for nonparametric regression. *Journal of the Royal Statistical Society B*, 76:71–96, 2014.

[6] John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[7] Thomas Sellke, M. J. Bayarri, and James Berger. Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55:62–71, 2001.

[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 Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society B*, 55:317–351, 1993.

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

[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 and Ivan Petej. Venn–Abers predictors, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 7, April 2014 (first posted in October 2012).

[15] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.