Training conformal predictors

Nicolo Colombo and **Vladimir Vovk** {NICOLO.COLOMBO,V.VOVK}@RHUL.AC.UK
Department of Computer Science, Royal Holloway University of London, Egham, Surrey, UK

Editor: Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov and Giovanni Cherubin

Abstract

Efficiency criteria for conformal prediction, such as *observed fuzziness* (i.e., the sum of p-values associated with false labels), are commonly used to *evaluate* the performance of given conformal predictors. Here, we investigate whether it is possible to exploit efficiency criteria to *learn* classifiers, both conformal predictors and point classifiers, by using such criteria as training objective functions.

The proposed idea is implemented for the problem of binary classification of handwritten digits. By choosing a 1-dimensional model class (with one real-valued free parameter), we can solve the optimization problems through an (approximate) exhaustive search over (a discrete version of) the parameter space. Our empirical results suggest that conformal predictors trained by minimizing their observed fuzziness perform better than conformal predictors trained in the traditional way by minimizing the *prediction error* of the corresponding point classifier. They also have reasonable performance in terms of their prediction error on the test set.

Keywords: Classification, criterion of efficiency, inductive conformal prediction, observed fuzziness.

1. Introduction

The standard approach to designing conformal predictors is to start from an existing machine-learning algorithm and turn it into a conformity measure (there may be more than one way of doing this). The rest is automatic: once we have a conformity measure, we can compute p-values, prediction sets, predictive distributions, etc. In this approach conformal prediction plays the role of a superstructure over traditional machine learning. The two parts of the resulting prediction algorithms, the traditional machine-learning part and the conformal part, are fairly autonomous and the interface between them is limited.

The standard approach has been fairly successful; conformal predictors have been built on top of a wide variety of traditional algorithms, including the Lasso (Lei, 2019), deep learning (Cortés-Ciriano and Bender, 2019), ridge regression, nearest neighbours, support vector machines, decision trees, and boosting (Vovk et al., 2005, Sections 2.3, 3.1, 4.2). However, the separation of the two parts may limit the power of this approach.

In this paper we propose blending the two parts of standard conformal prediction. The idea is to use existing criteria of efficiency for conformal prediction, such as those defined in Vovk et al. (2017). Instead of *evaluating* the performance of given conformal predictors, we propose to use those criteria for *training* conformal predictors by using such criteria as training objective functions.

We demonstrate the idea using binary classification of hand-written digits as example; see Section 3. As criterion of efficiency we use *observed fuzziness*, defined to be the sum of p-values for the test observations with their true labels replaced by false ones. This is one of the *probabilistic* criteria of efficiency; they are defined by Vovk et al. (2017), who argue that such criteria are akin to proper loss functions in machine learning and should be used in practice. The advantages of observed fuzziness over the other probabilistic criteria defined in Vovk et al. (2017) is that it does not depend on the significance level (which makes it easier to use) and does not include the noise created by the p-value for the true label (which makes it more stable).

For a simple 1-dimensional model class (i.e., involving one real-valued free parameters) we can solve the optimization problems used for training through an approximate exhaustive search over a discrete version of the parameter space. We compare two ways of training conformal predictors and point classifiers: using observed fuzziness and, as in traditional machine learning, using prediction error. In this context, the distinction between conformal predictors and point classifiers blurs: the former can be used as the latter (by using the label with the largest p-value as point prediction) and the latter, being defined in terms of a conformity measure, can be extended to the former. Our empirical results suggest, not surprisingly, that classifiers trained by minimizing observed fuzziness lead to a better observed fuzziness on the test set, and classifiers trained by minimizing prediction error lead to a better (but not overwhelmingly better) prediction error on the test set.

For computational efficiency, in this paper we concentrate on split-conformal prediction. Using full conformal prediction and other directions of further research are discussed in Section 4.

2. Background

The set of natural numbers is denoted $\mathbb{N} := \{1, 2, \dots\}$ (the positive integers). If a and b are two disjoint bags, we let $a + b := a \cup b$, and we use the notation $a + b$ only when the two (or more) bags are disjoint.

2.1. Observation space and data sets

Let $\mathcal{X}$ be a nonempty measurable *object space*, $\mathcal{Y}$ a discrete and finite *label space* of size $|\mathcal{Y}| \geq 2$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ the corresponding *observation space*. We will refer to elements of these sets as *objects*, *labels* and *observations*, respectively.

A *dataset* is a bag of elements of $\mathcal{Z}$, where a *bag* is a collection of elements (observations in this case) some of which may be identical (Vovk et al., 2005, Section 2.2). We will use the notation $\mathcal{Z}^{(*)}$ for the set of all datasets. Let us fix two nonempty datasets, the *training set* $\mathcal{D}_{\text{train}}$ and the *test set* $\mathcal{D}_{\text{test}}$.

In split-conformal prediction, the training set is randomly split into two disjoint bags, a *pre-training set* and a *pre-test set*, which we will denote as

$$\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{pre-train}} + \mathcal{D}_{\text{pre-test}}. \tag{1}$$

The pre-training set is often called the proper training set and the pre-test set is often called the calibration set (Vovk et al., 2005, Section 4.1), but our current terminology will be more convenient for this paper.

2.2. Conformity scores and p-values

A *conformity measure* is a function $Q : \mathcal{Z} \times \mathcal{Z}^{(*)} \to \mathbb{R}$ mapping each observation $(x, y) \in \mathcal{Z}$ and dataset $\mathcal{D}$ (such as the pre-training set $\mathcal{D}_{\text{pre-train}}$) to the corresponding *conformity score* $Q((x, y), \mathcal{D}) \in \mathbb{R}$. (A fuller term for Q would be "split-conformity measure". Our definition is adequate in this paper's context of split-conformal prediction, but in the context of full conformal prediction, in order to ensure validity we also need to require that Q should be invariant with respect to the permutations of the dataset.)

The *p-value* associated with an observation $(x, y) \in \mathcal{Z}$ (such as $(x, y) \in \mathcal{D}_{\text{test}}$), nonempty datasets $\mathcal{D}$ (such as $\mathcal{D}_{\text{pre-train}}$) and $\mathcal{D}'$ (such as $\mathcal{D}_{\text{pre-test}}$), and a conformity measure Q is

$$C(x, y, \mathcal{D}, \mathcal{D}', Q) := \frac{1 + \sum_{z \in \mathcal{D}'} \theta\left(Q((x, y), \mathcal{D}) - Q(z, \mathcal{D})\right)}{1 + |\mathcal{D}'|},$$

where the step function θ is defined by

$$\theta(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}, \quad u \in \mathbb{R}.$$

For $(x, y) \in \mathcal{D}_{\text{test}}$, $C(x, y, \mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, Q)$ shows how likely y is as the label of the test object x.

2.3. Observed fuzziness OF

The *observed fuzziness* of a conformity measure Q for nonempty datasets $\mathcal{D}$ (such as $\mathcal{D}_{\text{pre-train}}$), $\mathcal{D}'$ (such as $\mathcal{D}_{\text{pre-test}}$) and $\mathcal{D}''$ (such as $\mathcal{D}_{\text{test}}$) is

$$\text{OF}(\mathcal{D}, \mathcal{D}', \mathcal{D}'', Q) := \frac{\sum_{(x,y) \in \mathcal{D}''} \sum_{y' \in \mathcal{Y}} (1 - \delta_{y,y'}) C(x, y', \mathcal{D}, \mathcal{D}', Q)}{|\mathcal{D}''|}, \tag{2}$$

where $\delta_{y,y'}$ is defined by

$$\delta_{y,y'} = \begin{cases} 1 & \text{if } y = y' \\ 0 & \text{otherwise} \end{cases}, \quad y, y' \in \mathcal{Y}.$$

Note that one may choose $\mathcal{D}'' := \mathcal{D}'$; this may be useful at the stage of estimating a future observed fuzziness. For example, we may use $\text{OF}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, \mathcal{D}_{\text{pre-test}}, Q)$ to estimate $\text{OF}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, \mathcal{D}_{\text{test}}, Q)$ before seeing the test set. However, this introduces bias, since in this case each element of $\mathcal{D}''$ is also present in $\mathcal{D}'$, which is not expected to be the case for $\mathcal{D}'' := \mathcal{D}_{\text{test}}$. Therefore, we also define the three-argument version

$$\text{OF}(\mathcal{D}, \mathcal{D}', Q) := \frac{\sum_{(x,y) \in \mathcal{D}'} \sum_{y' \in \mathcal{Y}} (1 - \delta_{y,y'}) C(x, y', \mathcal{D}, \mathcal{D}' \setminus \{(x, y)\}, Q)}{|\mathcal{D}'|}$$

of (2).

2.4. Prediction error (PE)

The point predictor $\phi : \mathcal{X} \times \mathcal{Z}^{(*)} \to \mathcal{Y}$ obtained from a conformity measure Q and dataset $\mathcal{D}$ is defined as

$$\phi(x, \mathcal{D}, Q) \in \arg\max_{y \in \mathcal{Y}} Q((x, y), \mathcal{D});$$

we will assume that the arg max is a singleton (which is the case in our experiments). The *prediction error* of a conformity measure Q on nonempty datasets $\mathcal{D}$ (such as $\mathcal{D}_{\text{train}}$) and $\mathcal{D}'$ (such as $\mathcal{D}_{\text{test}}$) is defined as

$$\text{PE}(\mathcal{D}, \mathcal{D}', Q) := \frac{\sum_{(x,y) \in \mathcal{D}'} (1 - \delta_{y,y_*})}{|\mathcal{D}'|}, \quad y_* := \phi(x, \mathcal{D}, Q).$$

Note that, in this case, it is possible to use the entire training set as an input of the conformity measure, i.e., to let $\mathcal{D} := \mathcal{D}_{\text{train}}$ instead of $\mathcal{D} := \mathcal{D}_{\text{pre-train}}$.

3. Methods

For our experiments, in addition to the split (1), we consider a further split

$$\mathcal{D}_{\text{pre-train}} = \mathcal{D}_{\text{pre-pre-train}} + \mathcal{D}_{\text{pre-pre-test}}.$$

The overall split of the available data is

$$\mathcal{D}_{\text{train}} + \mathcal{D}_{\text{test}} = \mathcal{D}_{\text{pre-pre-train}} + \mathcal{D}_{\text{pre-pre-test}} + \mathcal{D}_{\text{pre-test}} + \mathcal{D}_{\text{test}}. \tag{3}$$

3.1. Model

We consider the binary classification problem of recognizing given hand-written digits in the MNIST dataset. We choose the conformity measure

$$Q_\rho((x, y), \mathcal{D}) := \frac{\sum_{(x',y') \in \mathcal{D}} \delta_{y,y'} \kappa_\rho(x, x')}{\sum_{(x',y') \in \mathcal{D}} \kappa_\rho(x, x')}, \quad \kappa_\rho(x, x') := e^{-\rho \|x - x'\|^2},$$

where $\rho \in \mathcal{R} \subseteq [0, \infty)$ is a free parameter. We use the datasets in the split (3) to train four models, Q_*^{PE}, $Q_*^{\text{pre-PE}}$, Q_*^{OF} and $Q_*^{\text{pre-OF}}$, defined by

$$\begin{aligned}
Q_*^{\text{PE}} &:= Q_{\rho_*}, & \rho_* &:= \arg\min_{\rho \in \mathcal{R}} \text{PE}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, Q_\rho), \\
Q_*^{\text{pre-PE}} &:= Q_{\rho_*}, & \rho_* &:= \arg\min_{\rho \in \mathcal{R}} \text{PE}(\mathcal{D}_{\text{pre-pre-train}}, \mathcal{D}_{\text{pre-pre-test}}, Q_\rho), \\
Q_*^{\text{OF}} &:= Q_{\rho_*}, & \rho_* &:= \arg\min_{\rho \in \mathcal{R}} \text{OF}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, Q_\rho), \\
Q_*^{\text{pre-OF}} &:= Q_{\rho_*}, & \rho_* &:= \arg\min_{\rho \in \mathcal{R}} \text{OF}(\mathcal{D}_{\text{pre-pre-train}}, \mathcal{D}_{\text{pre-pre-test}}, Q_\rho).
\end{aligned}$$

We evaluate the optimized models through the following performance scores:

$$\begin{aligned}
&\text{PE-train/PE-test} := \text{PE}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, Q_*^{\text{PE}}), \\
&\text{OF-train/PE-test} := \text{PE}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, Q_*^{\text{OF}}), \\
&\text{PE-train/OF-test} := \text{OF}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, \mathcal{D}_{\text{test}}, Q_*^{\text{pre-PE}}), \\
&\text{OF-train/OF-test} := \text{OF}(\mathcal{D}_{\text{pre-train}}, \mathcal{D}_{\text{pre-test}}, \mathcal{D}_{\text{test}}, Q_*^{\text{pre-OF}}).
\end{aligned}$$

k	PE-train/PE-test	OF-train/PE-test	PE-train/OF-test	OF-train/OF-test
0	0.190 ± 0.145	0.230 ± 0.155	0.402 ± 0.108	0.336 ± 0.052
1	0.120 ± 0.108	0.130 ± 0.110	0.367 ± 0.113	0.321 ± 0.060
2	0.180 ± 0.108	0.220 ± 0.154	0.386 ± 0.101	0.351 ± 0.066
3	0.210 ± 0.083	0.250 ± 0.112	0.432 ± 0.117	0.361 ± 0.075
4	0.210 ± 0.114	0.190 ± 0.114	0.410 ± 0.085	0.342 ± 0.070
5	0.150 ± 0.081	0.180 ± 0.098	0.445 ± 0.098	0.359 ± 0.103
6	0.150 ± 0.128	0.200 ± 0.134	0.407 ± 0.116	0.353 ± 0.047
7	0.200 ± 0.110	0.210 ± 0.114	0.426 ± 0.086	0.367 ± 0.051
8	0.260 ± 0.150	0.360 ± 0.102	0.417 ± 0.100	0.363 ± 0.107
9	0.230 ± 0.110	0.260 ± 0.150	0.449 ± 0.187	0.412 ± 0.089

Table 1: Case $n_{\text{size}} = 5$. The average and standard deviation of the scores obtained over 10 equivalent experiments with the sizes of the datasets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$. Integer $k \in \{0, \dots, 9\}$ in the first column is the digit to be discriminated in the corresponding experiments.

k	PE-train/PE-test	OF-train/PE-test	PE-train/OF-test	OF-train/OF-test
0	0.115 ± 0.059	0.115 ± 0.078	0.298 ± 0.040	0.272 ± 0.029
1	0.095 ± 0.072	0.125 ± 0.072	0.335 ± 0.062	0.299 ± 0.051
2	0.140 ± 0.058	0.150 ± 0.074	0.335 ± 0.042	0.315 ± 0.047
3	0.170 ± 0.117	0.205 ± 0.125	0.361 ± 0.092	0.300 ± 0.054
4	0.140 ± 0.070	0.220 ± 0.084	0.382 ± 0.057	0.325 ± 0.052
5	0.120 ± 0.068	0.110 ± 0.073	0.308 ± 0.050	0.331 ± 0.084
6	0.095 ± 0.057	0.170 ± 0.084	0.377 ± 0.103	0.288 ± 0.042
7	0.105 ± 0.069	0.115 ± 0.071	0.350 ± 0.086	0.309 ± 0.048
8	0.210 ± 0.089	0.260 ± 0.104	0.393 ± 0.076	0.333 ± 0.059
9	0.135 ± 0.084	0.195 ± 0.088	0.385 ± 0.068	0.293 ± 0.066

Table 2: Case $n_{\text{size}} = 10$. The average and standard deviation of the scores obtained over 10 equivalent experiments with the sizes of the datasets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$. Integer $k \in \{0, \dots, 9\}$ in the first column is the digit to be discriminated in the corresponding experiments.

Notice that for OF-testing we use the "pre-models" $Q_*^{\text{pre-PF}}$ and $Q_*^{\text{pre-OF}}$. This is because we want a valid conformal predictor; therefore, we do not touch the pre-test set at the stage of choosing our model. On the other hand, there is no need to worry about validity in the case of PE-testing; it is not guaranteed anyway.

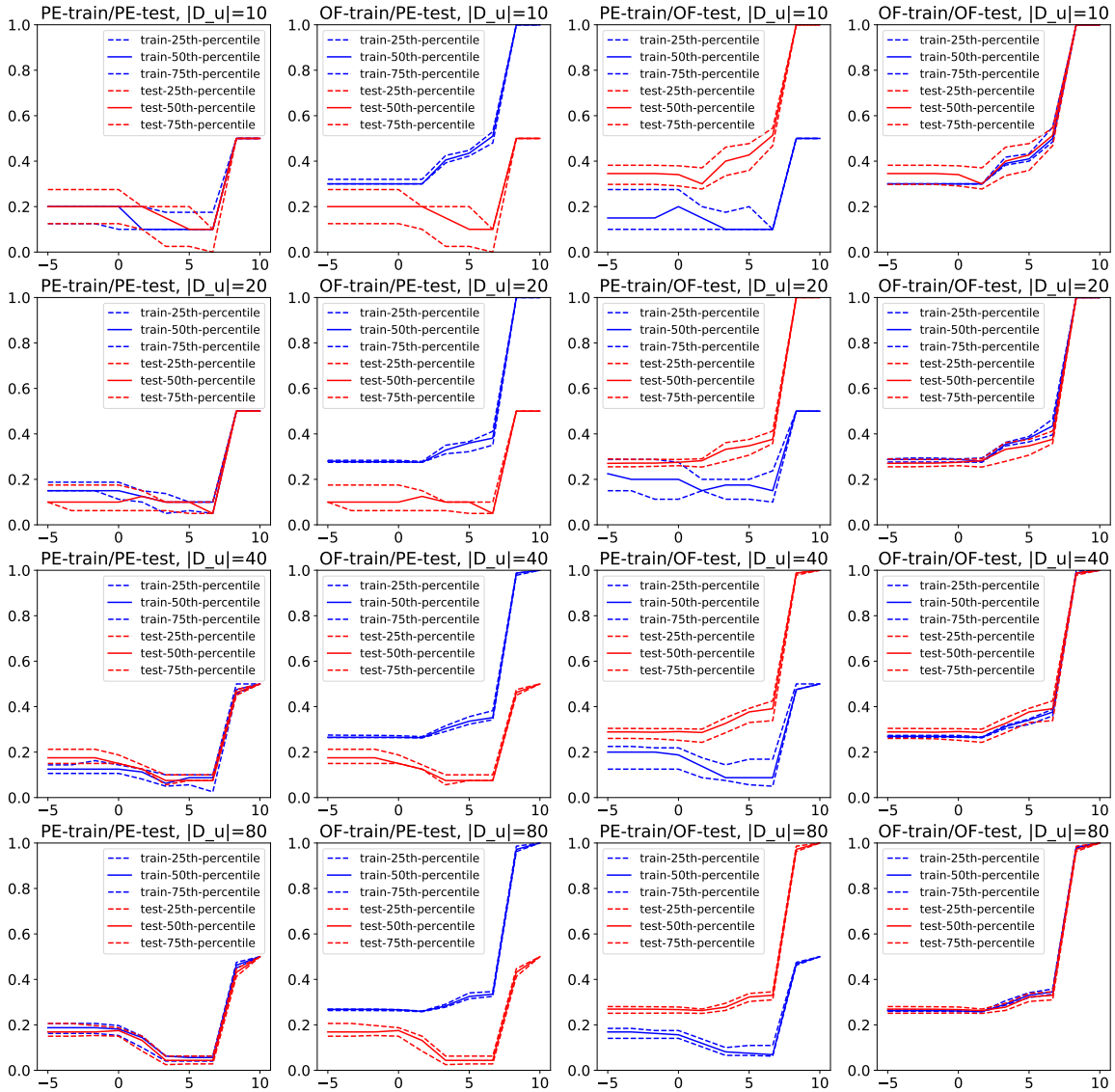

Figure 1: Digit $k = 0$. Values of the training and testing objective functions (PE or OF, y-axis) against the logarithm of the model free parameter ($\log \rho$, x-axis) for varying sizes of the datasets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$, $n_{\text{size}} \in \{5, 10, 20, 40\}$). Solid and dashed lines represent the median and the 25th or 75th percentile of the values obtained over 10 equivalent experiments. As specified in the plot legends, blue and red lines are associated with training and testing objectives, respectively. Plots obtained for $k > 0$ are similar.

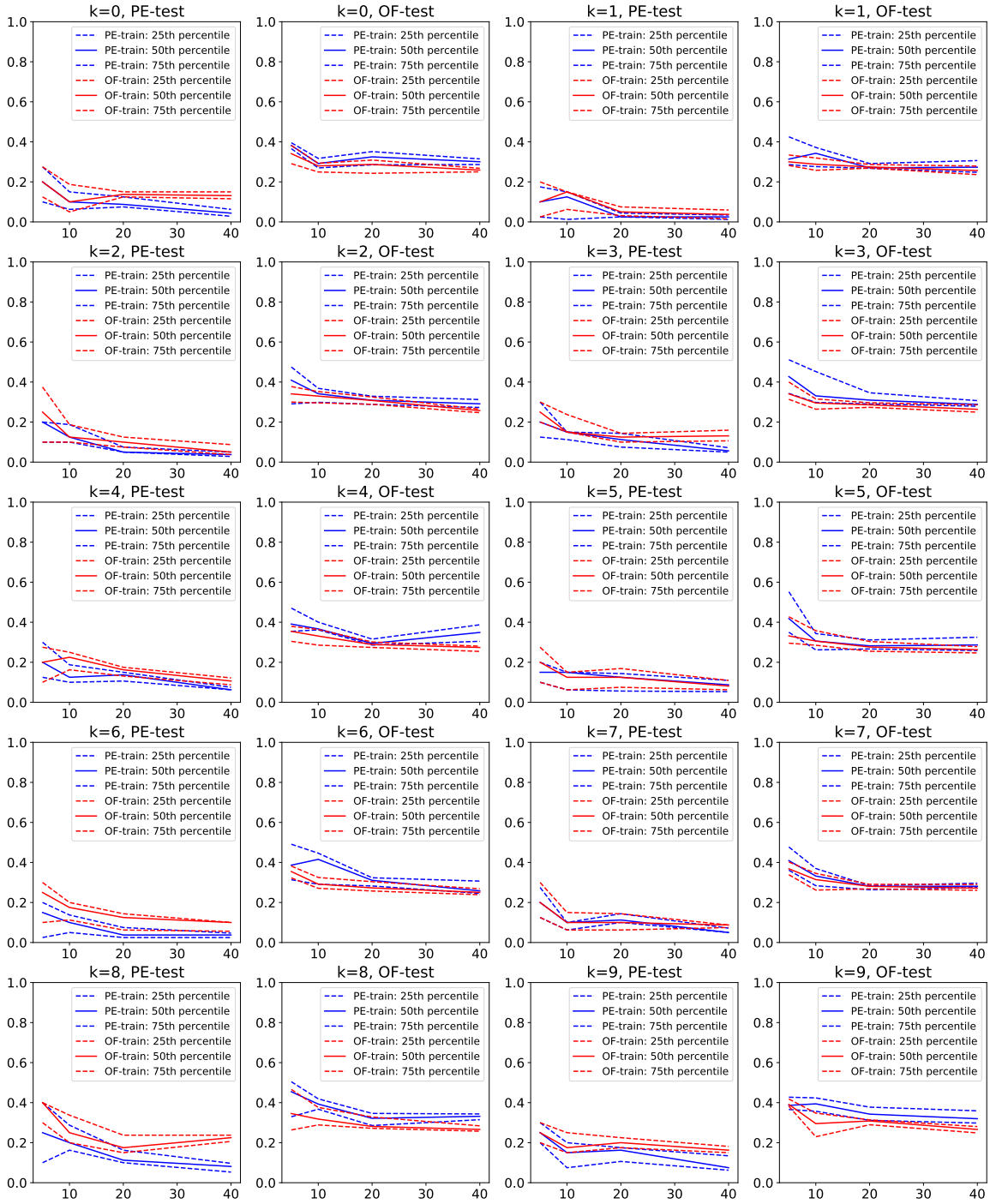

Figure 2: PE- and OF-test scores against the size of the training and test sets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$, $n_{\text{size}} \in \{5, 10, 20, 40\}$). Plot titles specify the integer $k \in \{0, \ldots, 9\}$ to be discriminated in the corresponding experiments and the type of testing objective. Each plot shows the median (solid lines) and the 25th or 75th percentile of the values obtained over 10 equivalent experiments. Blue and red lines are associated with PE- and OF-training, respectively.

k	PE-train/PE-test	OF-train/PE-test	PE-train/OF-test	OF-train/OF-test
0	0.097 ± 0.044	0.140 ± 0.032	0.320 ± 0.049	0.276 ± 0.037
1	0.035 ± 0.030	0.058 ± 0.043	0.280 ± 0.021	0.276 ± 0.020
2	0.063 ± 0.017	0.102 ± 0.039	0.310 ± 0.027	0.306 ± 0.024
3	0.125 ± 0.065	0.125 ± 0.034	0.316 ± 0.041	0.280 ± 0.020
4	0.130 ± 0.035	0.153 ± 0.048	0.303 ± 0.032	0.287 ± 0.022
5	0.115 ± 0.070	0.123 ± 0.070	0.291 ± 0.055	0.274 ± 0.038
6	0.053 ± 0.045	0.130 ± 0.076	0.307 ± 0.036	0.280 ± 0.028
7	0.120 ± 0.037	0.102 ± 0.061	0.280 ± 0.027	0.277 ± 0.022
8	0.125 ± 0.042	0.183 ± 0.058	0.325 ± 0.051	0.299 ± 0.041
9	0.143 ± 0.043	0.205 ± 0.063	0.343 ± 0.038	0.303 ± 0.021

Table 3: Case $n_{\text{size}} = 20$. The average and standard deviation of the scores obtained over 10 equivalent experiments with the sizes of the datasets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$. Integer $k \in \{0, \ldots, 9\}$ in the first column is the digit to be discriminated in the corresponding experiments.

k	PE-train/PE-test	OF-train/PE-test	PE-train/OF-test	OF-train/OF-test
0	0.049 ± 0.034	0.129 ± 0.040	0.303 ± 0.035	0.262 ± 0.022
1	0.029 ± 0.026	0.041 ± 0.026	0.279 ± 0.040	0.259 ± 0.027
2	0.045 ± 0.026	0.065 ± 0.030	0.291 ± 0.027	0.255 ± 0.016
3	0.069 ± 0.036	0.131 ± 0.030	0.290 ± 0.034	0.262 ± 0.026
4	0.076 ± 0.030	0.103 ± 0.027	0.345 ± 0.046	0.269 ± 0.017
5	0.082 ± 0.029	0.084 ± 0.030	0.295 ± 0.041	0.263 ± 0.020
6	0.036 ± 0.021	0.089 ± 0.047	0.275 ± 0.049	0.254 ± 0.025
7	0.055 ± 0.020	0.091 ± 0.024	0.291 ± 0.030	0.273 ± 0.016
8	0.078 ± 0.030	0.215 ± 0.043	0.330 ± 0.021	0.271 ± 0.020
9	0.092 ± 0.042	0.169 ± 0.029	0.327 ± 0.037	0.266 ± 0.023

Table 4: Case $n_{\text{size}} = 40$. The average and standard deviation of the scores obtained over 10 equivalent experiments with the sizes of the datasets $|\mathcal{D}_u| = 2n_{\text{size}}$ ($u \in \{\text{pre-pre-train}, \text{pre-pre-test}, \text{pre-test}, \text{test}\}$. Integer $k \in \{0, \ldots, 9\}$ in the first column is the digit to be discriminated in the corresponding experiments.

3.2. Experiments

We let $\rho \in \mathcal{R}$, with $\mathcal{R} := \{e^{\min_\rho + (\max_\rho - \min_\rho)r/R}\}_{r=0}^{R-1}$, $[\min_\rho, \max_\rho] := [-5, 10]$, $R := 10$. For given integers $k \in \{0, \dots, 9\}$ and dataset sizes $n_{\text{size}} \in \{5, 10, 20, 40\}$, we randomly extract from the MNIST database 10 quadruples of datasets listed on the right-hand side of (3) of equal sizes,

$$|\mathcal{D}_{\text{pre-pre-train}}| = |\mathcal{D}_{\text{pre-pre-test}}| = |\mathcal{D}_{\text{pre-test}}| = |\mathcal{D}_{\text{test}}| = 2n_{\text{size}}.$$

Each of the 400 datasets of size $2n_{\text{size}}$ contains n_{size} images of integer k (with label $y = 1$) and n_{size} images of other integers $k' \neq k$ (with label $y = 0$).

All objects $x \in \mathcal{X}$ are vectorized 28×28 grey-scale images of hand-written digits, i.e., all $x = (x_1, \dots, x_{784})$ ($x_i \in [0, 1]$, $i = 1, \dots, 784$) are such that the (i, j)th pixel of an image corresponds to the $(28(i-1)+j)$th entry of the associated vector x. The vectors are normalized to have the unit Euclidean length, $\|x\| = 1$ (this corresponds to normalizing the brightness of the images).

For each of the 400 binary classification datasets we run an independent training-testing experiment, as explained in Section 3.1. Summaries of results are reported in Figures 1–2 and Tables 1–4.

We have already commented on the results in Tables 1–4: when the goal is to design a good conformal predictor to be evaluated with the OF criterion, OF training is preferable (there is only one case in the tables where PE training works better); and for the design of point classifiers to be evaluated using prediction accuracy, PE training usually works better (there are only 3 cases where OF training works better and 2 cases where it works equally well). This finding can be summarized by saying that *consonant training* (OF-training when the goal is OF-performance and PE-training when the goal is PE-performance) usually works better than *dissonant training* (OF-training when the goal is PE-performance or PE-training when the goal is OF-performance). In our experiments, dissonant training works better in the case of OF-training (but of course, this needs to be confirmed on a wide range of datasets).

Figure 1 sheds light on the reasons for consonant training working better than dissonant training. In the case of consonant training (the first and fourth columns), the performance curves for training and test sets look very similar, and in many cases almost coincide. In the case of dissonant training (the second and third columns), they are not only at different levels (which is to be expected since the PE and OF criteria produce numbers at different scales), but their shapes look different, often attaining their minima at different places.

Figure 2 gives, essentially, a different representation of the results presented in Tables 1–4. The test performance improves as the size of the training set grows, albeit not very quickly.

4. Conclusion

In this paper we used a "validation set" (either the pre-test set or the pre-pre-test set) for choosing the conformity measure to use at the testing stage. There are ways to use the available training data more efficiently. On the PE side, we can use cross-validation. On the OF side, we can use cross-conformal prediction (or the related procedure of jack-

knife+, Barber et al., 2019) instead of split-conformal prediction, since the former are more economical with the data. This is an interesting direction of further research.

The performance of conformal predictors can be further improved (and provable validity of split-conformal prediction regained) by using full conformal prediction. However, in this case our methods will be computationally feasible only when applied to a fairly narrow (but important) class of training procedures.

Other directions of further research include:

- Replacing exhaustive search over the discretized parameter space used in this paper by more efficient methods, such as Gradient Descent.
- Experiments with other criteria of efficiency mentioned in Vovk et al. (2017), including those that are not probabilistic. (It is natural to expect that the prediction error for classifiers trained using such criteria suffers.)
- More extensive empirical studies covering a wide range of datasets.

Acknowledgements

Many thanks to Nicola Paoletti and Alex Gammerman for useful discussions. We are grateful to the three COPA 2020 reviewers for their thoughtful comments (some of which will become directions of further research for us). Vladimir Vovk's research has been supported by Amazon, Astra Zeneca, and Stena Line.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, December 2019.

Isidro Cortés-Ciriano and Andreas Bender. Deep confidence: A computationally efficient framework for calculating reliable prediction errors for deep neural networks. *Journal of Chemical Information and Modeling*, 59:1269–1281, 2019.

Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106:749–764, 2019.

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

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence*, 81:21–46, 2017.