Pattern Recognition 124 (2022) 108507

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Introduction to conformal predictors

Paolo Toccaceli

Department of Computer Science, Royal Holloway, University of London, UK

ARTICLE INFO

Article history:
Received 20 September 2021
Revised 18 December 2021
Accepted 22 December 2021
Available online 25 December 2021

2020 MSC:
68Q32
68T05

Keywords:
Conformal prediction
Nonparametric methods
Confidence

ABSTRACT

This paper aims to provide a compact but accessible introduction to Conformal Predictors (CP), a Machine Learning method with the distinguishing property of producing predictions that exhibit a chosen error rate. This property, referred to as validity, is backed by not only asymptotic, but also finite-sample probabilistic guarantees. CPs differ from the conventional approach to prediction in that they introduce hedging in the form of set-valued predictions. The CP validity guarantees do not require assumptions such as priors, but are of broad applicability as they rely solely on exchangeability. The CP framework is universal in the sense that it operates on top of virtually any Machine Learning method. In addition to the formal definition, this introduction discusses CP variants that can be computed efficiently (Inductive or “split” CP) or that are suitable for imbalanced data sets (class-conditional CP). Finally, a short survey of the field provides references for relevant research and highlights the variety of domains in which CPs have found valuable application.

1. Introduction

The most straightforward and arguably most common form of prediction consists in providing a single value, i.e. what we will refer to as a *bare prediction*. While this may be adequate in some simple applications, it is easy to convince ourselves that in fact it omits valuable information. The bare prediction does not convey the strength of the statistical evidence that supports its specific value; the strength of the evidence could be simply marginally lower for other values, but one would not be aware to this fact. Additional information should complement the bare prediction.

Conformal Predictors [1] convey this information in a novel way. They offer a principled, efficient, and flexible way to obtain predictions that guarantee a given error rate, under minimal assumptions. A distinctive feature of CP is that the predictions are hedged: they are sets, discrete in case of classification and continuous in the case of regression. As noted in Gammerman and Vovk [2], “the problem of hedged prediction is intimately connected with the problem of testing randomness”. Indeed, the theoretical foundations of CP can be traced to the universal test of randomness proposed by Per Martin-Löf [3].

CP have optimality properties. Quoting [4], we observe that “the predictions produced by the conformal algorithm are (a) invariant with respect to the [ordering of] old examples, (b) correct with the advertised probability, and (c) nested.[1] They are optimal among all region predictors with these properties”.

2. Notation

In the sequel, we will operate, unless otherwise stated, within the setting of *supervised learning*, using the follow notation and conventions.

A training set $\mathbf{Z} = \{z_1, z_2, \ldots, z_\ell\}$ contains *examples* $z_i = (x_i, y_i)$ consisting of an *object* $x_i \in \mathbf{X}$ and a *label* $y_i \in \mathbf{Y}$. We use the $\ell + 1$ index to denote the test object $x_{\ell+1}$.

We distinguish between two modes of operation, namely *batch* vs. *online*. In the batch mode, the training set[2] is provided once, in its entirety. The model is trained on it and then used to make an indefinite number of predictions. The actual labels of the test objects, possibly revealed after prediction, are not used to create new training examples. The order of the training examples and the test examples is irrelevant.

In the online mode, instead, the training examples are presented in a sequence. At step i, the ML method will have seen a training set $\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})\}$ and will be presented with a test object x_i. It will output a prediction $\hat{y}_i$ and the actual label

[1] The specific meaning of ‘nested’ is clarified in comments to Eq. (5).

[2] Note that in line with the established convention, the term *set* was used here, whereas in fact there is no stipulation for this collection not to contain multiple identical examples. Strictly speaking, the appropriate designation is training bag or training multiset and this will be used in Section 4.1 in line with the treatment in Vovk et al. [1].

E-mail address: Paolo.Toccaceli@rhul.ac.uk

https://doi.org/10.1016/j.patcog.2021.108507
0031-3203/

y_i will be revealed. The example $z_i = (x_i, y_i)$ will be added to the training set and the execution will move on to the next step $i + 1$.

3. Assumptions

The properties of CPs are guaranteed by a theoretical apparatus that rests on minimal assumptions. In this section, we will define such assumptions.

3.1. Independent and identically distributed random variables

Let's consider n random variables $V_1, \dots, V_n$ taking values in a space $\mathbf{V}$ and let F_{V_i} be the distribution of V_i. The $V_1, \dots, V_n$ variables are independent and identically distributed (i.i.d.) if and only if

$$F_{V_1}(v) = F_{V_k}(V) \qquad \forall k \in \{1, \dots, n\} \text{ and } \forall v \in \mathbf{V} \tag{1}$$

$$F_{V_1,\dots,V_n}(v_1, \dots, v_n) = F_{V_1}(v_1) \dots F_{V_n}(v_n) \qquad \forall v_1, \dots, v_n \in \mathbf{V} \tag{2}$$

This can be paraphrased by saying that the marginal probability of each variable is the same as that of any other of the n variables and that the joint distribution is simply the product of the marginal distributions for each variable. The i.i.d. assumption also means that the probability distribution of any Random Variable (RV) does not depend on the values of the other RVs. Informally, one could say that the knowledge of the values taken by the other RVs is of no help in predicting the distribution of any RV @. The definition above extends seamlessly to random vectors, which is in fact the case of interest in Machine Learning.

The i.i.d. assumption is pervasive in Machine Learning. The established setting of Statistical Learning Theory stipulates "random vectors $x \in R^n$ drawn independently from a fixed but unknown probability distribution $F(x)$" Vapnik [5, Page 17]. Indeed, the vast majority of ML algorithms relies, explicitly or implicitly, on the assumption that the test examples be drawn from the same distribution as the training examples. There are also, it has to be noted, approaches that do away even with this seemingly minimal requirement.[3]

3.2. Exchangeability

The variables $z_1, \dots, z_N$ are exchangeable if for every permutation τ of the integers $1, \dots, N$,

$$\Pr(z_1, \dots, z_n) = \Pr\big(z_{\tau(1)}, \dots, z_{\tau(n)}\big)$$

that is, the variables $w_1, \dots, w_N$, where $w_i = z_{\tau(i)}$, have the same joint probability distribution as $z_1, \dots, z_N$.

Exchangeability is effectively a property of the probability measure over the N random variables. To put it informally, the value of the probability measure does not depend on the order of its arguments.

It is straightforward to prove that i.i.d. variables are also exchangeable (the joint distribution is the product of N identical univariate distributions, hence the order is irrelevant).

The converse, however, is not true. An example of a sequence of RVs that is exchangeable but whose variables are not independent is Plya's urn (see, for instance, Lauritzen [7])

4. Conformal predictors

The framework of Conformal Prediction rests on the notion of Non-Conformity Measure (NCM). The NCM expresses how much it appears not to conform to a collection of samples. The NCM is the elementary tool we will use to assess randomness. Note that the NCM is not specified by the CP framework beyond the basic stipulation that it has to be a measurable function. The NCM is left to the user to supply. It is meant to be defined so that its values are larger, the more out-of-place the example appears. Fig. 1 illustrates the intuitive notion of Non-Conformity.[4]

The randomness of an example is then assessed by using the NCM in relative terms, rather than using its absolute value. To judge the randomness of an example, we determine the proportion of the examples in the collection that have a larger NCM than the example in hand. Low values of this proportion mean that it is rare to find examples that look more out-of-place, whereas high values signify that the majority of the examples would look more out-of-place.

The prediction set for a test object (of which we do not know the actual label) is built by the following procedure. For each possible value of the label, we construct a hypothetical example, made up of the test object and that hypothetical label (thereby forming a *hypothetical completion*), and we assess the randomness of these hypothetical completions. Only the labels for which the corresponding hypothetical completion exhibits a degree of randomness relative to the training set (measured as proportion of the training set with larger NCM) higher than the chosen significance level are included in the prediction set.

A key result proved in Vovk et al. [1, Theorem 8.1] establishes that, provided that data is exchangeable, a region predictor computed using this rule has an error rate that reflects the significance level (barring statistical fluctuation).

4.1. Formal definition

A rigorously formal definition of CP has inevitably to take into account a number of technicalities. Here, we will attempt to strike a balance between clarity and formality.

We will specify CP with reference to on-line mode as defined in Algorithm 1 because it is in this mode that the theoretical results are stated and proved in the literature, but it is possible to extend it in some sense also to the batch case.

```
Algorithm 1: On-line protocol.
  Data: Sequence of examples: $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, ...
  Result: Prediction sets: $\Gamma_{\epsilon,1}$, $\Gamma_{\epsilon,2}$, ...
          Cumulative error counts: $\mathrm{Err}_{\epsilon,1}, \mathrm{Err}_{\epsilon,2}, \dots$
  $\mathrm{Err}_0 = 0$ ;
  for $i = 1, 2, \dots$ do
      Nature presents $x_i$ ;
      Predictor outputs $\Gamma_{\epsilon,i}$ ;
      Nature reveals $y_i$ ;
      $\mathrm{err}_i = \begin{cases} 1 & \text{if} y_i \notin \Gamma_{\epsilon,i} \\ 0 & \text{otherwise} \end{cases}$ ;
      $\mathrm{Err}_{\epsilon,i} = \mathrm{Err}_{\epsilon,i-1} + \mathrm{err}_i$ ;
```

Let's assume that the training set is made up of a sequence of ℓ examples $z_i := (x_i, y_i) \in \mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ and $x_{\ell+1}$ is a test object taken from the same exchangeable distribution as the training examples.

We will use the notion of *bag* or *multi-set*. A bag of size $\ell \in \mathbb{N}$ is a collection of ℓ elements some of which may be identical; a bag differs from a set in that repetition is allowed. We will indicate a bag with the following notation $\{z_1, \dots, z_\ell\}$. The set of all possible

[3] One such example is the framework of Prediction with Expert Advice [6].

[4] CPs can be formulated equally well in terms of a Conformity Measure. The reason for using a possibly less intuitive Non-Conformity measure are explained in Section 4.8.

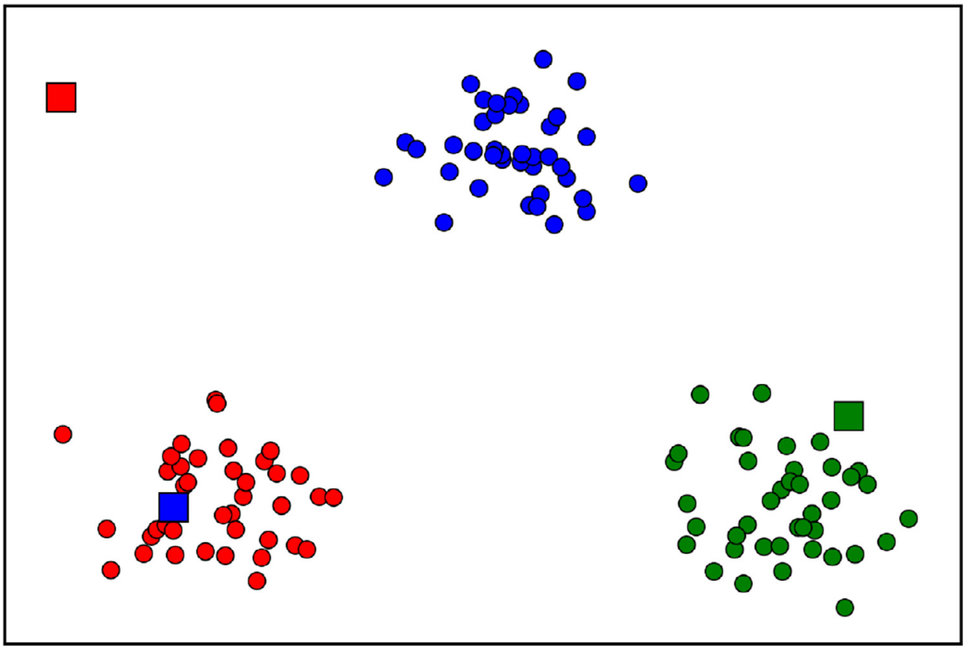

Fig. 1. An illustration of Non-Conformity. The round markers represent the collection of examples, where the colours red, green, blue correspond to the labels. The square markers represents test examples. The test example at the right (green square) does not look out of place, so a good Non-Conformity measure would assign a (relatively) low value to it, whereas the blue square in the middle of the red cluster at bottom left would have a high NCM. The case of the red marker at the top left is not as definite as those of the previous examples. The NCM would take an intermediate value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bags of k elements from a set $\mathbf{Z}$ will be denoted as $\mathbf{Z}^{(k)}$ (note the brackets around the exponent).

We call Non-Conformity Measure (NCM) a real-valued measurable function $A(z; \{z_1, \dots, z_k\}), A : \mathbf{Z} \times \mathbf{Z}^{(k)} \to \mathbb{R}$. The notation may seem strange, but it distinguishes clearly between the collection of examples $\{z_1, \dots, z_k\}$ and the example z for which we want to measure the non-conformity with respect to the collection. Also, the use of a bag emphasizes that the order of the elements is irrelevant and that the collection may contains identical elements.

Given a hypothetical value $\bar{y}_{\ell+1}$ for label for the test object, we form the so-called completion $z_{\ell+1} = (x_{\ell+1}, \bar{y}_{\ell+1})$. The NCM of each z_i is then computed according to:

$$\alpha_i = A((x_i, y_i), \{z_1, \dots, z_\ell, z_{\ell+1}\}/\{z_i\}) \qquad i = 1, \dots, \ell+1 \tag{3}$$

Given these NCM values, it is possible to compute for a p-value defined as:

$$p_{\bar{y}} := \frac{|\{i = 1, \dots, \ell+1 : \alpha_i \geq \alpha_{\ell+1}\}|}{\ell+1} \tag{4}$$

In words, the p-value of a hypothetical completion $(x_{\ell+1}, \bar{y}_{\ell+1})$ is the fraction of the elements in the training bag augmented with the hypothetical completion itself whose NCM is greater than or equal to the NCM of the hypothetical completion.[5]

The prediction set Γ_ϵ for a test object $x_{\ell+1}$ for a chosen significance level $\epsilon \in [0, 1]$ is the set of all the possible label values for which the p-value exceeds the significance level:

$$\Gamma_\epsilon(x) := \{y \mid p_y > \epsilon\} \tag{5}$$

Note that the prediction sets are nested in the sense that for $\epsilon_a < \epsilon_b < \dots < \epsilon_c$, it is the case that $\Gamma_{\epsilon_a} \supseteq \Gamma_{\epsilon_b} \supseteq \dots \supseteq \Gamma_{\epsilon_c}$ We say that an error occurs when the prediction set Γ_ϵ does not contain the actual label, i.e. $y_i \notin \Gamma_\epsilon$. We will refer to the count of the errors up to and including step n as $\mathrm{Err}_{\epsilon,n}$. It is customary to use the term *confidence* for the quantity $1 - \epsilon$.

As stated in previous sections, it is possible to state *validity* guarantees for CPs. With the definition of p-value in Eq. (4), it can be proved that the CP has a *conservative asymptotic validity* property in that the rate of errors converges almost surely[6] to a value that is less than or equal to the significance level, i.e.

$$\lim_{n\to\infty} \frac{\mathrm{Err}_{\epsilon,n}}{n} \leq \epsilon \qquad \text{a.s.} \tag{6}$$

To achieve exact validity, Eq. (4) must be modified so that ties (i.e. the occurrences of multiple α_i equal to $\alpha_{\ell+1}$) are broken with an element of randomness.

$$p_{\bar{y}} := \frac{|\{i = 1, \dots, \ell+1 : \alpha_i > \alpha_{\ell+1}\}| + \tau |\{i = 1, \dots, \ell+1 : \alpha_i = \alpha_{\ell+1}\}|}{\ell+1} \tag{7}$$

where $\tau \sim U[0, 1]$, i.e. τ is an RV uniformly distributed on $[0, 1]$ (this RV is to be "drawn" independently for each test object). With this more complex definition of p-value (referred to as *smoothed* p-value), it can be proved that the asymptotic validity becomes exact, i.e.

$$\lim_{n\to\infty} \frac{\mathrm{Err}_{\epsilon,n}}{n} = \epsilon \qquad \text{a.s.} \tag{8}$$

The validity property can be formally proved as a consequence of the following key theorem Gammerman and Vovk [2, Theorem 1]

Theorem 1. *Suppose the examples $(x_1, y_1), (x_2, y_2), \dots$ are generated independently from the same probability distribution.*

For any smoothed Conformal Predictor working in the on-line prediction protocol and any significance level $\epsilon \in [0, 1]$, the Random Vari-

[5] An equivalent formulation might have used the bag unchanged and simply added one at the numerator. This formulation however will become preferable when we introduce the smoothed conformal predictor.

[6] The 'almost surely' qualification is a technicality that can be informally explained as the fact that the probability of encountering a sequence of examples such that the assertion is not verified is vanishing.

ables $\mathrm{err}_1, \mathrm{err}_2, \ldots,$ *are independent and take value 1 with probability* ϵ.

Both conservative and exact validity guarantees as stated above in Eqs. (6) and (8) are asymptotic and it may be argued that they may not be relevant in any finite-sample regime that we may encounter in practice. There exists also a finite-sample guarantee Vovk et al. [1, p.27], which can be derived by applying Hoeffding's inequality:

$$\forall n > 0,\ \forall \delta > 0 \qquad \mathbb{P}\left[\frac{\mathrm{Err}_{\epsilon,n}}{n} \geq (\epsilon + \delta)\right] \leq e^{-2n\delta^2} \tag{9}$$

In words, this finite-sample guarantee states that for any choice of $\delta > 0$, the probability that the actual observed error rate exceeds the targeted ϵ by δ is bounded by $e^{-2n\delta^2}$.

4.2. *CP for regression*

The definition of CP in the previous section appears to suggest that a prediction set for a test object is computed by examining as many cases as possible values of the label. Obviously, this would be infeasible for regression problems because in that setting the label can take infinitely many values. However, for some definitions of NCMs, it turns out that it is possible to compute (theoretically exact) prediction sets in a finite number of steps. Two observations are relevant here: (a) there is a finite number of possible values that the p-value can take and (b) what we need to compute is really what values of the label correspond to a given p-value. The former is a direct consequence of the definition itself of CP p-value. The latter is a direct consequence of the definition of prediction set as the set of labels for which the p-value is greater than the significance level. For some choices of NCM, the relationship between p-values and hypothetical labels can be computed in a finite number of steps. A basic standard approach is described in Papadopoulos et al. [9] whereas Nouretdinov et al. [8] proposes an efficient method based on Ridge Regression (later extended to a kernel-based form). A promising development is that of Conformal Predictive Distributions [10], a method of computing predictive distributions without requiring priors, in a frequentist setting, with validity guarantees.

4.3. *Efficiency*

The prediction sets $\Gamma_\epsilon(x)$ can contain any subset of the label space $\mathbf{Y}$, including the entire label space $\mathbf{Y}$ and the empty set. The latter happens when there is no label $\bar{y}_{\ell+1}$ that would create with the test object $x_{\ell+1}$ an example whose non-conformity (with respect to the examples in the bag) would be small enough for the example to be included in the prediction. One way to interpret this is that the object is an anomaly,[7] in that no label assignment would "seem right". An empty prediction is automatically counted as a prediction error.

It has to be noted that conservative validity, in the sense of a guarantee that the predictions will exhibit an error rate that does not exceed the chosen significance level ϵ, can be banally obtained by predicting always the entire set $\mathbf{Y}$ as $\Gamma_\epsilon(x)$. Of course, such predictions would be totally uninformative and completely useless. In fact we want the prediction sets $\Gamma_\epsilon(x)$ to be as small as possible (without being empty)[8] We can therefore identify two main desiderata in set prediction:

- Validity: the error rate corresponds to the chosen significance level.
- Efficiency: the prediction sets are as small as possible

There is obviously a trade-off in general between these two goals, as making the prediction sets smaller makes missing the correct label more likely. By using CPs, one can take advantage of the fact that validity is guaranteed, so that all efforts can be focused solely on improving efficiency.[9]

Note that validity is guaranteed regardless of the choice of NCM @. Even if the NCM is a constant, the smoothed CP exhibits exact validity; however the predictions are uninformative. It is the efficiency, i.e. the extent of the hedging, that is determined by the NCM @. A Non-Conformity Measure can be in principle extracted from any Machine Learning (ML) algorithm, which is then referred to as the *underlying* ML method. Although there is no universal method to derive an NCM, a default choice is:

$$A((x, y), \{z_1, \ldots, z_k\}) := \Delta(y, f(x)) \tag{10}$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is the prediction rule learned on $(z_1, \ldots, z_k)$ and $\Delta : \mathbf{Y} \times \mathbf{Y}' \to \mathbb{R}$ is a measure of dissimilarity between a label and a prediction. For a survey of NCMs used in published CP applications, the interested reader can refer to Section 11.3 in Balasubramanian et al. [12]. Note also that any monotone transformation of an NCM produces the CP that outputs the same predictions as the original CP (by looking at Eq. (4) the p-values are exactly the same).

4.4. *Transductive and inductive CP*

CP as described in the previous section is referred to as Transductive. The term originates from an idea put forward by Vapnik [5, p.293]. Stated very briefly, the idea transcends the conventional inductive approach by proposing to exploit the knowledge of the test object $x_{\ell+1}$ when training. Indeed, when computing the NCM, a hypothetical example with the test object and a hypothetical label is added to the bag in Eq. (3). The underlying ML model is retrained for each α_i, $i = 1, \ldots, \ell + 1$, $\{z_1, \ldots, z_\ell, z_{\ell+1}\}/(x_i, y_i)$, that is, a training set in which the ith example has been removed. Limiting ourselves to the classification case, in which the set $\mathbf{Y}$ is discrete, for one test object, the underlying ML model is trained $|\mathbf{Y}| \cdot (\ell + 1)$ times. In practice, the resulting computational cost becomes prohibitive for all but the simplest practical applications.

A different form of CP has been proposed [9] which prescribes a different way of computing the NCM and retains the validity property. This different form is referred to as Inductive CP or, by some authors more recently (see, for instance, Lei et al. [13]), as Split CP @. As illustrated schematically in Fig. 2, the training set is partitioned into two sets, called proper training set and calibration set.[10] The proper training set is used to train the underlying ML method. The training of the underlying ML method is performed once only. The same fitted model is used to compute the α_i for the examples of the calibration set and for the $\alpha_{\ell+1}$ on the hypothetical examples formed by trying out in turn every possible label.

[7] One should keep in mind that, as discussed later in Section 4.8, the p-values for the correct labels are distributed uniformly. So, empty prediction sets could occur also, with probability that depends on the significance level, when the object is not an anomaly.

[8] We argue that empty prediction sets are not desirable in the specific case in which we are dealing with objects that have a label and we want to have a prediction that identifies that label. However, one can conceive other scenarios, in which empty prediction sets might convey a useful result (e.g. anomaly detection).

[9] This echoes a direction advocated by the eminent statistician Tukey [11].

[10] It could be argued that Inductive CPs require more data because they need a calibration set in addition to a proper training set. While intuitively justifiable, there is no theoretical result (as far as the author was able to ascertain) that prescribes the partition into calibration and training set. Cross-conformal predictors [14] can mitigate this issue but the averaging that is recommended in the paper results in loss of validity.

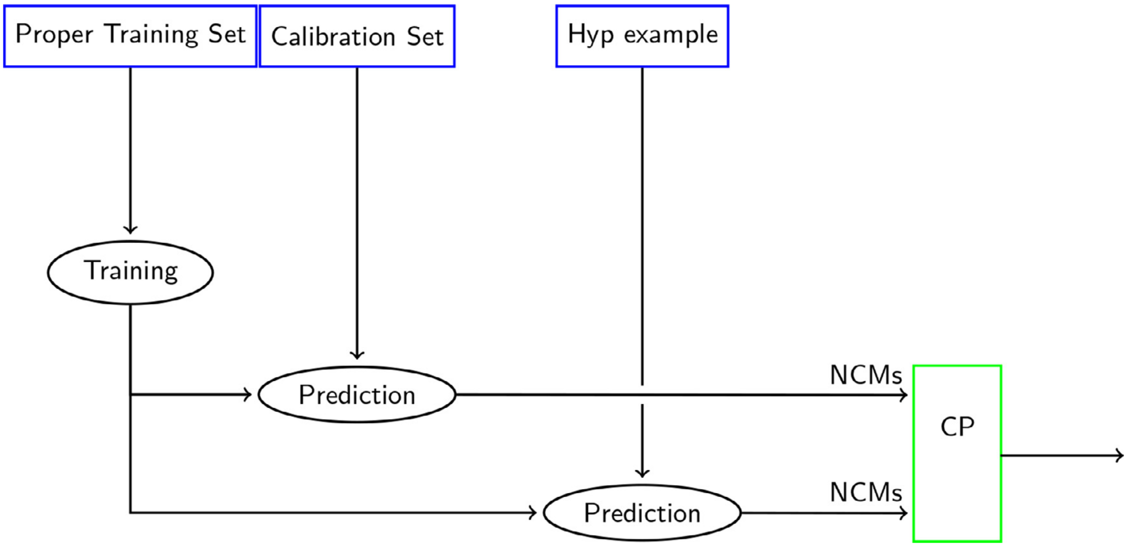

Fig. 2. A schematic representation of Inductive CP.

Assuming that the first m examples constitute the calibration set and the remaining $k = \ell - m$ examples the proper training set, the α_i can be formally expressed as:

$$\alpha_i = A((x_i, y_i), \{z_1, \dots, z_k\}) \qquad i = 1, \dots, m$$
$$\alpha_{\ell+1} = A((x_{\ell+1}, y_{\ell+1}), \{z_1, \dots, z_k\})$$

The (smoothed) p-value for a hypothesis $y_{\ell+1} = \bar{y}$ about the label of test object $x_{\ell+1}$ is defined as follows:

$$p_{\bar{y}} = \frac{|\{i = 1, \dots, m+1 : \alpha_i > \alpha_{m+1}\}| + \tau |\{i = 1, \dots, m+1 : \alpha_i = \alpha_{m+1}\}|}{m+1} \tag{11}$$

It is interesting to note that, for Inductive CP, the assumption of exchangeability is required only for the calibration set and the test set. The validity property is in fact independent of the proper training set, which can then be chosen arbitrarily. In fact, it can also be observed that there is no inherent reason for the NCM to have to be learned on a proper training set. It can be chosen at will, possibly based on some a priori knowledge, as long as it satisfies the requirements stated in Section 4.1.

In terms of algorithmic asymptotic complexity, it is worthwhile noting that, omitting some details, an efficient implementation of the p-value calculation might first order the calibration set NCMs α_i with one-off computational cost $O(n \log n)$ where n is the cardinality of the calibration set and then for each test example $\alpha_{\ell+1}$ determine the index in the resulting ordered vector, which can be done with cost $O(\log n)$. This would compute the left term at the numerator in Eq. (11); the right term is obtained with similar considerations.

4.5. *Label-conditional (Mondrian) CP*

Finally, it is important to note that the validity property as stated above guarantees the chosen error rate as long as the test examples are i.i.d. with respect to the calibration examples. One consequence of this is that there is in general no validity guarantee if we compute the relative frequency of errors only over test examples of a chosen label, i.e. we consider label-conditional error rates. However, the per-label validity can be achieved with a variant of CP, called *label-conditional CP* (or also Mondrian[11] CP). The label-conditional CP is actually one form of more general conditional CPs, which are discussed in broader terms in Vovk [15]. Label-conditional CPs differ only in the calculation of the p-value: we restrict the α_i only to those that are associated with examples with the same label as the hypothetical label that we are assigning at the test object. So, the p-value for a hypothesis $y_{\ell+1} = \bar{y}$ about the label of test object $x_{\ell+1}$ is defined as follows:

$$p(\bar{y}) = \frac{|\{i = 1, \dots, (m+1) : y_i = \bar{y}, \alpha_i \ge \alpha_{m+1}\}|}{|\{i = 1, \dots, (m+1) : y_i = \bar{y}\}|} \tag{12}$$

The property of label-conditional validity is essential in practice when the CP is applied to an "imbalanced" data set, that is, a data set in which the proportions of labels are significantly different. Empirically, one can observe that with the plain validity property, the overall error rates tend within statistical fluctuation to the chosen significance level, but the minority class(es) are disproportionally affected by errors (see, for instance, Löfström et al. [16]). This property ensures that, even for the minority class, the long-term error rate will tend to the chosen significance level.

4.6. *An example on synthetic data*

In this section, we will walk through an example of Transductive CP on a binary classification problem using k Nearest Neighbours as the underlying ML method. A synthetic training data set is generated so that the examples form two roughly semicircular interlocking clouds of points in the plane (commonly referred to as "moons"), as illustrated in Fig. 3.

The NCM is chosen as:

$$\alpha := \frac{\sum^{(k)}_{j \neq i : y_j = y_i} d(x_j, x_i)}{\sum^{(k)}_{j \neq i : y_j \neq y_i} d(x_j, x_i)}$$

[11] Conditional CPs are formally defined by introducing a notion of taxonomy on the space of the examples, on the basis of which the training set (or the calibration set in the Inductive CP case) is partitioned into categories. The graphical representation of this partitioning on bivariate examples gives rise to images that can remind one of the distinctive style associated with the Dutch–French artist Piet Mondrian.

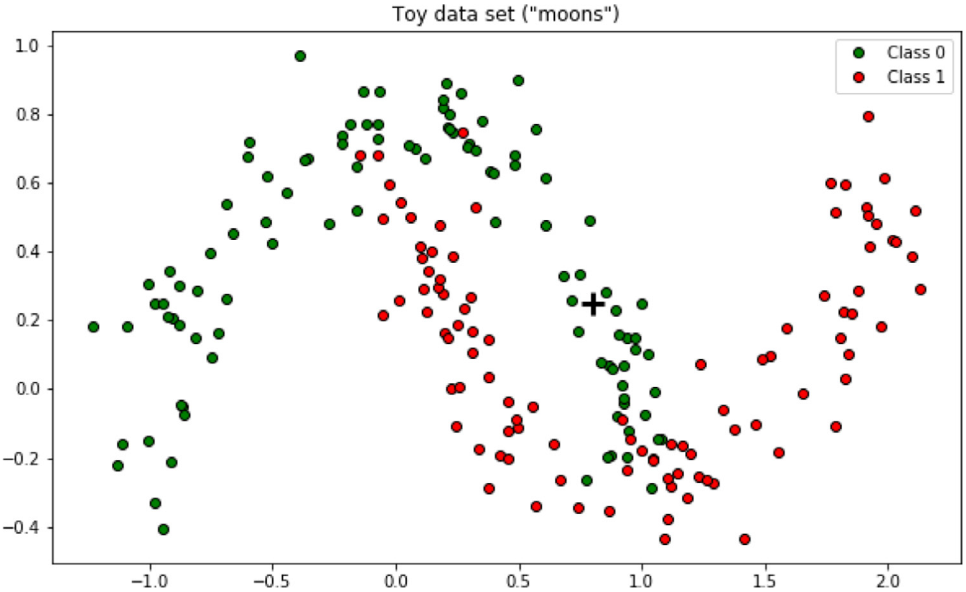

Fig. 3. A synthetic data set (moons) for the binary classification example. For the test object represented by the black cross at coordinates (0.8,0.25), the p-values are $p_0 = 0.5622$ and $p_1 = 0.00995$. If we chose $\epsilon = 0.01$, the prediction set would be {0}, label 1 being rejected because $p_1 < 0.01$.

where by $\sum^{(k)}$ we denote the sum of only the k smallest terms. In this example, k was set to 3. So, given a training set $\{z_1, \ldots, z_\ell\}$ and a hypothetical test example $z_{\ell+1}$, the ith α (with $i = 1, \ldots, \ell = 1$) is calculated by removing z_i from the bag $\{z_1, \ldots, z_{\ell+1}\}$ and calculating the ratio between the sum of the distances of the $k = 3$ closest examples with the same label as z_i and the sum of the $k = 3$ closest examples with a different label than z_i.

For any test point there are four possibilities as to the prediction:

	Prediction set
$p_0 \le \epsilon,\ p_1 \le \epsilon$	$\varnothing$
$p_0 > \epsilon,\ p_1 \le \epsilon$	$\{0\}$
$p_0 \le \epsilon,\ p_1 > \epsilon$	$\{1\}$
$p_0 > \epsilon,\ p_1 > \epsilon$	$\{0, 1\}$

Fig. 4 shows the predictions for every test object in a rectangular region, using a different colour for each of the four possible outcomes listed above. As the target error rate ϵ is reduced, the areas where the CP makes single predictions shrink. The CP outputs more uncertain predictions (i.e. predictions sets containing more than one value) as it cannot reject at that significance level any of the labels.

One basic technique to assess the performance of classification models is to use a confusion matrix, i.e. a form of contingency table in which rows correspond to the actual labels of test examples, columns to the predicted labels, and the cells (i, j) contain the counts of examples of label i predicted as j.

In the case of CP, the confusion matrix takes a slightly different form than usual as a consequence of the different nature of the predictions, which are sets rather than single values. Considering only the binary classification case, for each actual label, we may be interested in how many examples were correctly predicted (with the prediction set containing only the correct label), how many examples were incorrectly predicted (that is, the prediction set contains only the incorrect label) how many examples were predicted inconclusively (that is, the prediction set contains both labels), and finally how many examples were given an empty prediction set.

A summary of the predictions for different significance levels is shown in Table 1. The error rate is within statistical fluctuation of the significance level, consistently with the validity property of CP @. The data set in this example is imbalanced, with one third of the examples of class 1 and the remaining two thirds of class 0.

As discussed in Section 4.5, when the data set is imbalanced, the validity guarantee of CP does not apply to each class separately. Indeed, one can observe that in Table 1 the error rate for class 1 is markedly greater than the significance level. To remedy this, one can use label-conditional CP @. In Table 2, label-conditional CP is applied to the same training and test data as in Table 1. The resulting error rate for class 0 and error rate for class 1 are both close to the significance level, as graphically illustrated in Fig. 5.

4.7. Confidence and credibility

Restricting now our attention to classification problems, in some cases, it may be desirable to focus the hedged forecast on a single value referred to as point prediction, rather than a set or an interval. The most straightforward choice is to take the label that is associated with the largest p-value[12] The hedging of the prediction can then be expressed by complementing the point prediction with quantities that characterize the uncertainty. For example, Saunders et al. [17] and Gammerman and Vovk [2] recommend using *confidence* and *credibility*. Confidence is defined as:

$$\sup\{1 - \epsilon \ : |\Gamma_\epsilon| \le 1\}$$

that is, the largest "confidence" (in sense defined in Section 4.1) for which the prediction set contains only one label. It can be computed as 1-second largest p_y.

Credibility is:

$$\inf\{\epsilon : |\Gamma_\epsilon| = 0\}$$

which can also be expressed more simply as the largest p-value and, as such, it is also the lower bound for the value of ϵ that would result in an empty prediction.

It has to be noted that there are no theoretical guarantees on these two quantities.

[12] This is not necessarily equivalent to taking the highest scoring label according to the underlying ML method. In the case of Mondrian CP, for instance, the p-value for label $\bar{y}$ is calculated with respect to the calibration set examples with label $\bar{y}$.

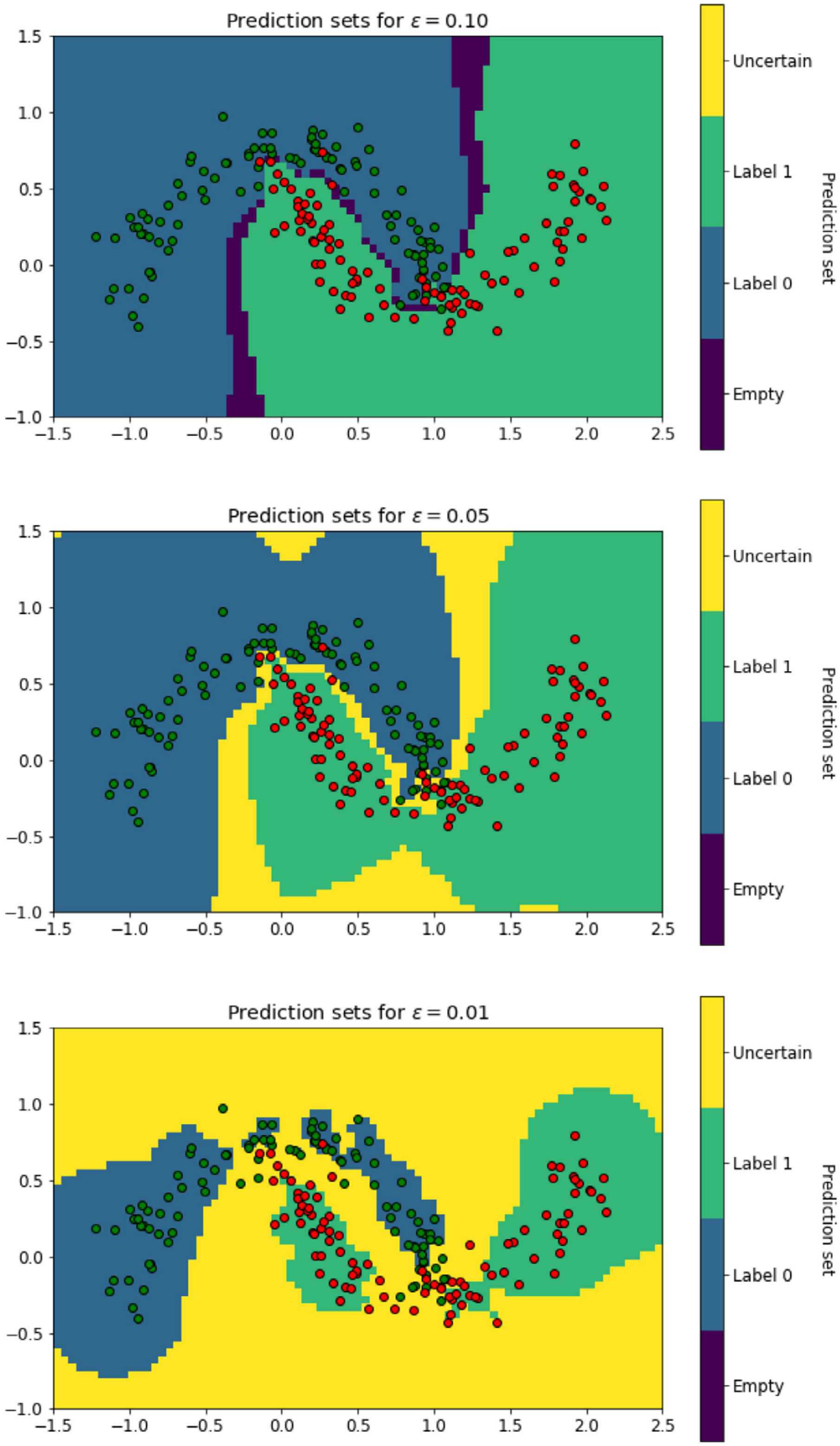

Fig. 4. Each plot shows the prediction sets at a given significance level for a grid of test points. The colour at each point codes whether the prediction set for the test object at that point is empty (dark blue), contains only label 0 (light blue), contains only label 1 (green), or contains both labels and hence is an uncertain prediction (yellow). As the significance level is decreased (i.e. we demand a lower error rate) the blue areas shrink and eventually yellow areas (where the prediction sets contain both labels) take their place. The eventual prevalence of the yellow areas arises because the label hypotheses can no longer be rejected at such low significance levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Example of CP predictions for various values of the significance level ϵ. The error rate is reported next to the significance level ϵ to facilitate the verification of the validity property. By the comparison of the two columns, one can confirm that the validity property does hold, within statistical fluctuation. The entries of the confusion matrices are presented (arranged linearly) in the central group of 8 columns. The test data set had 1200 class 0 examples and 400 class 1 examples.

ϵ	Error rate	Uncertain fraction	1 pred {1}	1 pred {0}	0 pred {0}	0 pred {1}	1 pred $\varnothing$	0 pred $\varnothing$	1 pred {0, 1}	0 pred {0, 1}	1 error rate	0 error rate
0.05	0.056	0.011	329	62	1163	28	0	0	9	9	0.155	0.023
0.10	0.111	0.000	310	33	1113	9	57	78	0	0	0.225	0.072
0.15	0.165	0.000	294	15	1042	1	91	157	0	0	0.265	0.132
0.20	0.204	0.000	281	8	992	0	111	208	0	0	0.297	0.173
0.25	0.251	0.000	259	4	939	0	137	261	0	0	0.352	0.217
0.50	0.497	0.000	109	0	695	0	291	505	0	0	0.728	0.421
0.75	0.744	0.000	27	0	383	0	373	817	0	0	0.932	0.681
0.80	0.783	0.000	21	0	326	0	379	874	0	0	0.948	0.728
0.85	0.829	0.000	14	0	259	0	386	941	0	0	0.965	0.784
0.90	0.892	0.000	7	0	166	0	393	1034	0	0	0.983	0.862
0.95	0.943	0.000	1	0	91	0	399	1109	0	0	0.998	0.924

Table 2

Example of label-conditional CP predictions for various values of the significance level ϵ. This uses the same data set (in fact, the same α_i) as Fig. 1, but computes the p-values using the label-conditional method of Eq. (12). The validity property holds at the level of each label, as the two rightmost columns show.

ϵ	Error rate	Uncertain fraction	1 pred {1}	1 pred {0}	0 pred {0}	0 pred {1}	1 pred $\varnothing$	0 pred $\varnothing$	1 pred {0, 1}	0 pred {0, 1}	1 error rate	0 error rate
0.05	0.043	0.094	359	11	1021	58	0	0	30	121	0.028	0.048
0.10	0.106	0.000	364	19	1067	72	17	61	0	0	0.090	0.111
0.15	0.160	0.000	335	9	1009	35	56	156	0	0	0.163	0.159
0.20	0.222	0.000	308	4	937	8	88	255	0	0	0.230	0.219
0.25	0.284	0.000	291	2	855	1	107	344	0	0	0.273	0.287
0.50	0.507	0.000	225	0	564	0	175	636	0	0	0.438	0.530
0.75	0.719	0.000	123	0	326	0	277	874	0	0	0.693	0.728
0.80	0.768	0.000	84	0	287	0	316	913	0	0	0.790	0.761
0.85	0.830	0.000	69	0	203	0	331	997	0	0	0.828	0.831
0.90	0.893	0.000	44	0	127	0	356	1073	0	0	0.890	0.894
0.95	0.958	0.000	16	0	52	0	384	1148	0	0	0.960	0.957

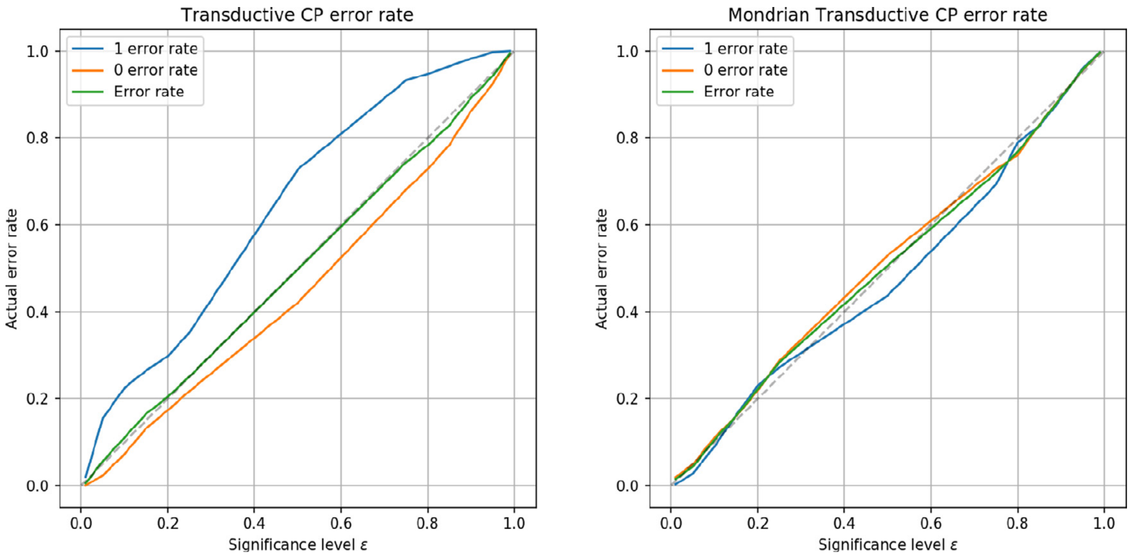

Fig. 5. Label-conditional validity. The plot on the left shows that the plain CP exhibits validity overall (green line), but deviates significantly from it on the minority class. This issue does not occur in the plot on the right, where label-conditional CP is used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.8. CP and statistical hypothesis testing

The framework of CP can be interpreted as an application of the methods of "traditional" Statistical Hypothesis Testing (SHT) to Machine Learning. Indeed the p-value can be viewed as the probability of drawing from the same distribution **F** that generated the training set an example that is as or more contrary to the hypothesis of randomness than the one in hand. The prediction set for an object $x_{\ell+1}$ is then formed by all the labels $\bar{y}$ for which the Null Hypothesis that the (hypothetical) example $(x_{\ell+1}, \bar{y})$ comes from **F** cannot be rejected at the chosen significance level ϵ. The NCM plays the role of test statistic, i.e. of a value that is larger the more contrary to the Null Hypothesis a sample is. This explains why, instead of a Conformity Measure, the possibly less straightforward notion of Non-Conformity Measure is used in defining @.

In the light of these considerations, one can see how the p-value defined with Eq. (7) can indeed be viewed as cognate of the p-value in SHT. It is then important to warn against its misinterpretation as the posterior probability of the Null Hypothesis. Despite receiving ample treatment in the statistical literature (for example, Sellke et al. [18]) such "p-value fallacy", as it is often referred to, remains still insidious and widespread.

It is worth noting that Vovk et al. [1, Ch. 6] proposed another class of predictors, under the name of Venn Predictors, which operate on top of generic classification methods and output probabilistic predictions. Dispensing with some technicalities, it can be stated that those probabilities are calibrated, in the sense that they reflect long-term frequencies (that is, the relative frequency of label a among the examples for which the VP outputs $p_a = k$ is k, within statistical fluctuation).

5. Survey of the field

Conformal Prediction can be viewed as part of the long tradition of nonparametric estimation [19,20]. The earliest published exposition of the concepts dates back to 1998 [21]. The reference text is the monograph "Algorithmic Learning in a Random World" published in 2005 [1], which presents in a coherent whole the fundamental theoretical results gradually established by the authors in the previous decade.

Among the papers that followed, Gammerman and Vovk [2] summarized succinctly the key aspects of CP and was complemented by a very interesting discussion in which leading scholars provided their perspectives. In more recent times, Balasubramanian et al. [12] collected contributions from key researchers in the field and provided a thorough overview of the state-of-the-art as of 2015 for both CP theory and CP applications. To avoid repetitions of material and references covered in that book, this section will focus on the developments that occurred after 2014.

An increasing number of academic groups around the world appear to be involved in research on CP. The one with the longest tradition is arguably the Centre for Reliable Machine Learning,[13] at Royal Holloway, University of London, where the founders of the field of Conformal Prediction, Prof. Alexander Gammerman and Prof. Vladimir Vovk, continue to drive forward the research, developing new ideas, such as Conformal Predictive Distributions [10] and Conformal Test Martingales [22].

A significant body of research with acknowledged connections to CP have been made by scholars at Carnegie-Mellon University, often in collaboration with researchers from the University of Chicago and Stanford University. Among the many contributions, we highlight a method to "conformalize" the LASSO [23], a method to compensate for a form of deviation at test time from the i.i.d. assumption (namely, covariate shift) [24] (which was recently extended to distribution shift in Gibbs and Candès [25]), a way to exploit CP to make deep learning image classifiers more robust [26], a modification of Cross-Conformal Predictors [27], a method to apply CP to quantile regression algorithms[14] algorithms [28] and investigations on the limits of distribution-free inference [29].

Another prolific group of researchers operates in Sweden, gravitating around the KTH Royal Institute of Technology in Stockholm, the Stockholm University, the Jonkoping University, Uppsala University, and the University of Boras, often in connection with the Discovery Science department of the pharmaceutical company AstraZeneca. Restricting our survey to the last 5 years, the studies that community contributed focus on combination of CP [30], on

[13] https://cml.rhul.ac.uk/, known as Computer Learning Research Centre (CLRC) up to 2019.

[14] Quantile regression refers to conditional quantile functions, where quantile function appear to be another name for predictive distribution.

the use of random forests as underlying ML method [31], on interpretable conformal regression [32], and on several applications listed further down.

Other centres of intense research activity include Fredrik University in Cyprus (under the guidance of Harris Papadopoulos, who made significant contributions) and Maastricht University in the Netherlands (Evgueni Smirnov).

The main forum for the CP researchers is the yearly Symposium on "Conformal and Probabilistic Prediction with Applications" (COPA) which started in 2012. The papers presented at the symposium are published as Proceedings of Machine Learning Research.[15] The latest at the time of writing is PMLR Vol. 128 (COPA2020), whereas for earlier editions see for instance [33]. CP has also been the focus of conferences, such as the 2015 "DST-EPSRC Indo-UK Workshop on Conformal Prediction and Applications" in Hyderabad and the 2015 "Statistical Learning and Data Sciences" Symposium in Egham, U @.

A number of journals have featured special issues on CP @. These include: Annals of Mathematics and Artificial Intelligence [34], Journal of Cheminformatics [35], Machine Learning [36], Neurocomputing [37].

In recent years, CP has seen a plethora of applications. A large number of those can be grouped under the banner of life sciences. CP has been applied in neuropsychology to predict the progress of Alzheimer's Disease [38,39], in a biomedical setting to predict lung cancer survival [40] or breast cancer survivability [41] or detecting seizures [42] or detecting lung cancer using a electronic nose [43] in ecology to predict aquatic toxicity [44] but perhaps the lion's share of applications is in chemoinformatics [35] and, more specifically, drug discovery [45–48] and development, where CP is used in high-throughput screening [49–53], toxicity studies [54–56], animal testing alternatives [57], proteochemometric studies [58].

There are in addition a variety of applications in surprisingly disparate fields, including nuclear fusion [59], identification of maggots in forensics [60], malware detection [61–63], fairness in the justice system [64], recommender systems [65,66], detection of anomalous trajectories [67], classification of herbal medicines with an electronic nose [68], predictive maintenance [69] and predictive monitoring of Hybrid Automata[16] [70].

Applications of CP are not confined to research settings. Several companies are known to employ CP techniques in software used in "production" systems. Among them are AstraZeneca in Sweden [71] and Janssen in Belgium [72].

6. Summary and conclusions

Conformal Predictors offer a theoretically sound framework for obtaining set-valued predictions exhibiting a chosen error rate. CPs can operate in an on-line or in a batch mode and in a classification as well as in a regression setting, under the minimal assumption of exchangeability. They do not require to postulate priors and do not rely on any assumption as to the form of the probability distributions of objects and labels. They are universal in the sense that they can be applied on top of virtually any ML method. They can be efficiently computed which makes them applicable to large data sets. Finally, about two decades of research have given rise a number of variants that have found application in several disparate fields (e.g. chemoinformatics).

This article aimed at providing a succinct but gentle introduction to the main underlying ideas, with formal proofs and further details deliberately left to the references provided. It is hoped that this accessible guide will give researchers and practitioners the information they need to make CP a valuable addition to their ML toolbox.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author is grateful to Prof. Alexander Gammerman for his invaluable advice and patient guidance.

References

[1] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[2] A. Gammerman, V. Vovk, Hedging predictions in machine learning (with discussion), Comput. J. 50 (2) (2007) 151–163, doi:10.1093/comjnl/bxl065.

[3] P. Martin-Lf, The definition of random sequences, Inf. Control 9 (6) (1966) 602–619, doi:10.1016/S0019-9958(66)80018-9.

[4] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (2008) 371–421.

[5] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, Heidelberg, 1995.

[6] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.

[7] S. Lauritzen, Exchangeability and de Finetti's Theorem, 2007. www.stats.ox.ac.uk

[8] I. Nouretdinov, T. Melluish, V. Vovk, Ridge regression confidence machine, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 385–392.

[9] H. Papadopoulos, K. Proedrou, V. Vovk, A. Gammerman, Inductive Confidence Machines for Regression, vol. 2430, 2002. 10.1007/3-540-36755-1_29

[10] V. Vovk, J. Shen, V. Manokhin, M.-G. Xie, Nonparametric predictive distributions based on conformal prediction, Mach. Learn. 108 (3) (2019) 445–474, doi:10.1007/s10994-018-5755-8.

[11] J.W. Tukey, Sunset salvo, Am. Stat. 40 (1) (1986) 72–76, doi:10.1080/00031305.1986.10475361.

[12] V. Balasubramanian, S. Ho, V. Vovk, Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications, Elsevier Science, 2014.

[13] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, J. Am. Stat. Assoc. 113 (523) (2018) 1094–1111, doi:10.1080/01621459.2017.1307116.

[14] V. Vovk, Cross-conformal predictors, Ann. Math. Artif. Intell. 74 (1) (2015) 9–28, doi:10.1007/s10472-013-9368-4.

[15] V. Vovk, Conditional validity of inductive conformal predictors, Mach. Learn. 92 (2–3) (2013) 349–376, doi:10.1007/s10994-013-5355-6.

[16] T. Löfström, H. Boström, H. Linusson, U. Johansson, Bias reduction through conditional conformal prediction, Intell. Data Anal. 19 (6) (2015) 1355–1375, doi:10.3233/IDA-150786.

[17] C. Saunders, A. Gammerman, V. Vovk, Transduction with confidence and credibility, in: IJCAI International Joint Conference on Artificial Intelligence, vol. 2, 1999, pp. 722–726.

[18] T. Sellke, M.J. Bayarri, J.O. Berger, Calibration of p values for testing precise null hypotheses, Am. Stat. 55 (1) (2001) 62–71, doi:10.1198/000313001300339950.

[19] L. Wasserman, All of Nonparametric Statistics, Springer Texts in Statistics, Springer New York, 2010.

[20] A.B. Tsybakov, Introduction to Nonparametric Estimation, first ed., Springer Publishing Company, Incorporated, 2008.

[21] A. Gammerman, V. Vapnik, V. Vovk, Learning by transduction, in: Proceedings of the Fourteenth Conference on Uncertainty in Articial Intelligence, Morgan Kaufmann, 1998, pp. 148–156.

[22] V. Vovk, I. Petej, I. Nouretdinov, E. Ahlberg, L. Carlsson, A. Gammerman, Retrain or not retrain: conformal test martingales for change-point detection, 2021.

[23] J. Lei, Fast exact conformalization of the lasso using piecewise linear homotopy, Biometrika 106 (4) (2019) 749–764, doi:10.1093/biomet/asz046.

[24] R.J. Tibshirani, R.F. Barber, E. Candes, A. Ramdas, Conformal prediction under covariate shift, in: Advances in Neural Information Processing Systems, 2019, pp. 2526–2536.

[25] I. Gibbs, E. Candès, Adaptive Conformal Inference Under Distribution Shift, (2021) arXiv e-prints arXiv:2106.00170

[26] Y. Hechtlinger, B. Póczos, L. Wasserman, Cautious deep learning, arXiv preprint arXiv:1805.09460(2018).

[15] http://proceedings.mlr.press/

[16] Hybrid Automata can be seen mathematical descriptions of devices that combine digital logic with analog processes (e.g. a cardiac pacemaker). They can be described as a finite state machine with a finite set of continuous variables, possibly governed by a system of differential equations.

[27] R.F. Barber, E.J. Candes, A. Ramdas, R.J. Tibshirani, Predictive inference with the jackknife+(2019).

[28] Y. Romano, E. Patterson, E.J. Candès, Conformalized quantile regression(2019).

[29] R.F. Barber, Is distribution-free inference possible for binary regression? (2020).

[30] H. Linusson, U. Johansson, H. Boström, Efficient conformal predictor ensembles, Neurocomputing (2019), doi:10.1016/j.neucom.2019.07.113.

[31] T. Vasiloudis, G. de Francisci Morales, H. Boström, Quantifying uncertainty in online regression forests, J. Mach. Learn. Res. 155 (2019) 1–35.

[32] U. Johansson, C. Sonstrod, T. Lofstrom, H. Bostrom, Customized interpretable conformal regressors, in: Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019, 2019, pp. 221–230, doi:10.1109/DSAA.2019.00037.

[33] A. Gammerman, Z. Luo, J. Vega, V. Vovk, Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20–22, 2016, Proceedings, in: Lecture Notes in Computer Science, Springer International Publishing, 2016.

[34] A. Gammerman, V. Vovk, Foreword to this special issue: conformal and probabilistic prediction with applications, Ann. Math. Artif. Intell. 81 (1) (2017) 1–2, doi:10.1007/s10472-017-9557-7.

[35] O. Spjuth, Novel applications of machine learning in cheminformatics, J. Cheminform. 10 (1) (2018) 46, doi:10.1186/s13321-018-0301-z.

[36] A. Gammerman, V. Vovk, H. Boström, L. Carlsson, Conformal and probabilistic prediction with applications: editorial, Mach. Learn. 108 (3) (2019) 379–380, doi:10.1007/s10994-018-5761-x.

[37] A. Gammerman, V. Vovk, Z. Luo, E. Smirnov, R. Peeters, Special issue on conformal and probabilistic prediction with applications, Neurocomputing (2019), doi:10.1016/j.neucom.2019.11.025.

[38] T. Pereira, S. Cardoso, D. Silva, A. de Mendonça, M. Guerreiro, S.C. Madeira, Towards trustworthy predictions of conversion from mild cognitive impairment to dementia: a conformal prediction approach, in: International Conference on Practical Applications of Computational Biology & Bioinformatics, Springer, 2017, pp. 155–163.

[39] T. Pereira, S. Cardoso, M. Guerreiro, S.C. Madeira, A.D.N. Initiative, et al., Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, venn-ABERS, and conformal predictors: a case study in AD, J. Biomed. Inform. 101 (2020) 103350.

[40] K. Qaddoum, Lung cancer Patient's Survival Prediction Using GRNN-CP, in: Communications in Computer and Information Science, vol. 1187, CCIS, 2020, pp. 143–150, doi:10.1007/978-3-030-43364-2_13.

[41] L.M. Alnemer, L. Rajab, I. Aljarah, Conformal prediction technique to predict breast cancer survivability, Int. J. Adv. Sci. Technol. 96 (2016) 1–10.

[42] C. Eliades, H. Papadopoulos, Detecting seizures in EEG recordings using conformal prediction, in: Conformal and Probabilistic Prediction and Applications, 2018, pp. 171–186.

[43] X. Zhan, Z. Wang, M. Yang, Z. Luo, Y. Wang, G. Li, An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction, Measurement 158 (2020), doi:10.1016/j.measurement.2020.107588.

[44] F. Svensson, U. Norinder, Conformal prediction for ecotoxicology and implications for regulatory decision-making, in: Methods in Pharmacology and Toxicology, 2020, pp. 271–287, doi:10.1007/978-1-0716-0150-1_12.

[45] M. Eklund, U. Norinder, S. Boyer, L. Carlsson, The application of conformal prediction to the drug discovery process, Ann. Math. Artif. Intell. 74 (1–2) (2015) 117–132.

[46] E. Ahlberg, O. Hammar, C. Bendtsen, L. Carlsson, Current application of conformal prediction in drug discovery, Ann. Math. Artif. Intell. 81 (1–2) (2017) 145–154.

[47] I. Cortés-Ciriano, A. Bender, Concepts and applications of conformal prediction in computational drug discovery, in: Artificial Intelligence in Drug Discovery, The Royal Society of Chemistry, 2021, pp. 63–101, doi:10.1039/9781788016841-00063.

[48] N. Bosc, F. Atkinson, E. Felix, A. Gaulton, A. Hersey, A.R. Leach, Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery, J. Cheminform. 11 (1) (2019) 4.

[49] P. Toccaceli, I. Nouretdinov, A. Gammerman, Conformal predictors for compound activity prediction, in: A. Gammerman, Z. Luo, J. Vega, V. Vovk (Eds.), Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20–22, 2016, Proceedings, Springer International Publishing, Cham, 2016, pp. 51–66, doi:10.1007/978-3-319-33395-3_4.

[50] J. Sun, L. Carlsson, E. Ahlberg, U. Norinder, O. Engkvist, H. Chen, Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets, J. Chem. Inf. Model. 57 (7) (2017) 1591–1598, doi:10.1021/acs.jcim.7b00159.

[51] F. Svensson, N. Aniceto, U. Norinder, I. Cortes-Ciriano, O. Spjuth, L. Carlsson, A. Bender, Conformal regression for quantitative structure-activity relationship modeling—Quantifying prediction uncertainty, J. Chem. Inf. Model. 58 (5) (2018) 1132–1140, doi:10.1021/acs.jcim.8b00054.

[52] L. Ahmed, V. Georgiev, M. Capuccini, S. Toor, W. Schaal, E. Laure, O. Spjuth, Efficient iterative virtual screening with Apache Spark and conformal prediction, J. Cheminform. 10 (1) (2018), doi:10.1186/s13321-018-0265-z.

[53] F. Svensson, A.M. Afzal, U. Norinder, A. Bender, Maximizing gain in high-throughput screening using conformal prediction, J. Cheminform. 10 (1) (2018) 7, doi:10.1186/s13321-018-0260-4.

[54] F. Svensson, U. Norinder, A. Bender, Modelling compound cytotoxicity using conformal prediction and PubChem HTS data, Toxicol. Res. 6 (1) (2017) 73–80.

[55] C. Ji, F. Svensson, A. Zoufir, A. Bender, eMolTox: prediction of molecular toxicity with confidence, Bioinformatics (Oxford, England) 34 (14) (2018) 2508–2509, doi:10.1093/bioinformatics/bty135.

[56] A. Morger, M. Mathea, J.H. Achenbach, A. Wolf, R. Buesen, K.-J. Schleifer, R. Landsiedel, A. Volkamer, KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development, J. Cheminform. 12 (1) (2020) 24, doi:10.1186/s13321-020-00422-x.

[57] A. Forreryd, U. Norinder, T. Lindberg, M. Lindstedt, Predicting skin sensitizers with confidence—Using conformal prediction to determine applicability domain of GARD, Toxicol. Vitro 48 (2018) 179–187, doi:10.1016/j.tiv.2018.01.021.

[58] I. Cortés-Ciriano, A. Bender, T. Malliavin, Prediction of PARP inhibition with proteochemometric modelling and conformal prediction, Mol. Inform. 34 (6–7) (2015) 357–366.

[59] R. Moreno, J. Vega, S. Dormido, J. Contributors, Conformal prediction of disruptions from scratch: application to an ITER scenario, in: Symposium on Conformal and Probabilistic Prediction with Applications, Springer, 2016, pp. 67–74.

[60] S. Beyramysoltan, M.I. Ventura, J.Y. Rosati, J.E. Giffen-Lemieux, R.A. Musah, Identification of the species constituents of maggot populations feeding on decomposing remains—Facilitation of the determination of post mortem interval and time since tissue infestation through application of machine learning and direct analysis in real time-mass spectrometry, Anal. Chem. 92 (7) (2020) 5439–5446, doi:10.1021/acs.analchem.0c00199.

[61] G. Cherubin, I. Nouretdinov, A. Gammerman, R. Jordaney, Z. Wang, D. Papini, L. Cavallaro, Conformal clustering and its application to botnet traffic, in: A. Gammerman, V. Vovk, H. Papadopoulos (Eds.), Statistical Learning and Data Sciences, Springer International Publishing, Cham, 2015, pp. 313–322.

[62] S.K. Dash, G. Suarez-Tangil, S. Khan, K. Tam, M. Ahmadi, J. Kinder, L. Cavallaro, Droidscribe: classifying android malware based on runtime behavior, in: 2016 IEEE Security and Privacy Workshops (SPW), IEEE, 2016, pp. 252–261.

[63] W. Zhi, H.-z. Gao, Y.-m. Zhang, Y.-c. Hu, K.-f. Qiu, X. Cheng, C.-f. Jia, Fortifying botnet classification based on venn-abers prediction, DEStech Transactions on Computer Science and Engineering (CST), 2017.

[64] Y. Romano, R.F. Barber, C. Sabatti, E.J. Candès, With malice towards none: assessing uncertainty via equalized coverage (2019).

[65] S. Ayyaz, U. Qamar, R. Nawaz, HCF-CRS: a hybrid content based fuzzy conformal recommender system for providing recommendations with confidence, PLoS One 13 (10) (2018) e0204849, doi:10.1371/journal.pone.0204849.

[66] T. Himabindu, V. Padmanabhan, A. Pujari, Conformal matrix factorization based recommender system, Inf. Sci. 467 (2018) 685–707, doi:10.1016/j.ins.2018.04.004.

[67] R. Laxhammar, G. Falkman, Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories, Ann. Math. Artif. Intell. 74 (1–2) (2015) 67–94, doi:10.1007/s10472-013-9381-7.

[68] X. Zhan, X. Guan, R. Wu, Z. Wang, Y. Wang, Z. Luo, G. Li, Online conformal prediction for classifying different types of herbal medicines with electronic nose, in: IET Conference Publications, vol. 2018, 2018, doi:10.1049/cp.2018.1730.

[69] I. Nouretdinov, J. Gammerman, M. Fontana, D. Rehal, Multi-level conformal clustering: adistribution-free technique for clustering and anomaly detection, Neurocomputing (2019), doi:10.1016/j.neucom.2019.07.114.

[70] L. Bortolussi, F. Cairoli, N. Paoletti, S. Smolka, S. Stoller, Neural Predictive Monitoring, in: Lecture Notes in Computer Science, vol. 11757, LNCS, 2019, pp. 129–147, doi:10.1007/978-3-030-32079-9_8.

[71] H. Chen, T. Kogej, O. Engkvist, Cheminformatics in drug discovery, an industrial perspective, Mol. Inform. 37 (9–10) (2018) 1800041.

[72] K. Kumar, V. Chupakhin, A. Vos, D. Morrison, D. Rassokhin, M.J. Dellwo, K. McCormick, E. Paternoster, H. Ceulemans, R.L. DesJarlais, Development and implementation of an enterprise-wide predictive model for early absorption, distribution, metabolism and excretion properties, Future Med. Chem. 13 (19) (2021) 1639–1654, doi:10.4155/fmc-2021-0138.

Paolo Toccaceli got his Electronic Engineering M.Sc. from Politecnico di Milano in 1993 and worked in various technical roles in the R&D of high-tech companies such as HP and Alcatel-Lucent. In 2021 he got a Ph.D. in Machine Learning at Royal Holloway, Univ. of London. He now works at Graphcore.