Conformal predictive decision making

Vladimir Vovk and Claus Bendtsen

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 #19

First posted July 22, 2017. Last revised December 23, 2018.

Project web site:
http://alrw.net

Abstract

This note explains how conformal predictive distributions can be used for the purpose of decision-making. Namely, a major limitation of conformal predictive distributions is that, at this time, they are only applicable to regression problems, where the label is a real number; however, this does not prevent them from being used in a general problem of decision making. The resulting methodology of conformal predictive decision making is illustrated on a small benchmark data set. Our main theoretical observation is that there exists an asymptotically efficient predictive decision-making system which can be obtained by using our methodology (and therefore, satisfying the standard property of validity).

Contents

1 Introduction

Conformal predictive distributions were introduced in Vovk et al. (2017) as result of combining the notion of predictive distributions (Schweder and Hjort, 2016, Chapter 12, and Shen et al., 2018) with the method of conformal prediction (Vovk et al., 2005). This note shows how they can be applied in decision making. Since conformal prediction satisfies the standard property of validity, our methods will lead to efficient decisions when applied to prediction algorithms with good resolution (even if they are miscalibrated).

Let $\mathbf{X}$ and $\mathbf{Y}$ be fixed non-empty measurable spaces; we will call them the *object* and *label spaces*, respectively. The Cartesian product $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is the *observation space*. Suppose we are given a training sequence $z_1, \dots, z_n$ of observations $z_i = (x_i, y_i) \in \mathbf{Z}$. In this note we discuss the following decision problem: given an object $x \in \mathbf{X}$ and a set $\mathbf{D}$ of available decisions, choose the best decision $d \in \mathbf{D}$ for that object. A major limitation of conformal predictive distributions, defined in Section 2, is that at this time they only solve the problem of probabilistic regression, where labels are real numbers. The purpose of this note is to explain that they are still applicable to our decision problem.

2 Conformal predictive distributions

In this section we define conformal predictive distributions following Vovk et al. (2017). We let $U_{0,1}$ stand for the uniform probability measure on the interval $[0,1]$ and consider the real line $\mathbb{R}$ as the label space. A function $Q : (\mathbf{X} \times \mathbb{R})^{n+1} \times [0,1] \to [0,1]$, where $n \in \mathbb{N} := \{1, 2, \dots\}$ is a natural number, is called a *randomized predictive system* if:

R1a For each training sequence $(z_1, \dots, z_n) \in (\mathbf{X} \times \mathbb{R})^n$ and each test object $x \in \mathbf{X}$, the function $Q(z_1, \dots, z_n, (x, y), \tau)$ is monotonically increasing both in y and in τ.

R1b For each training sequence $(z_1, \dots, z_n) \in (\mathbf{X} \times \mathbb{R})^n$ and each test object $x \in \mathbf{X}$,

$$\begin{aligned} \lim_{y \to -\infty} Q(z_1, \dots, z_n, (x, y), 0) &= 0, \\ \lim_{y \to \infty} Q(z_1, \dots, z_n, (x, y), 1) &= 1. \end{aligned} \tag{1}$$

R2 For any probability measure P on $\mathbf{X} \times \mathbb{R}$, the distribution of Q (as function of random training observations $z_1 \sim P, \dots, z_n \sim P$, a random test observation $z \sim P$, and a random number $\tau \sim U_{0,1}$, all assumed independent) is uniform:

$$\forall \alpha \in [0,1] : \mathbb{P}\{Q(z_1, \dots, z_n, z, \tau) \le \alpha\} = \alpha.$$

A *conformity measure* is a measurable function $A : (\mathbf{X} \times \mathbb{R})^{n+1} \to \mathbb{R}$ that is invariant with respect to permutations of the first n observations: for any

sequence $(z_1, \dots, z_n) \in (\mathbf{X} \times \mathbb{R})^n$, any $z \in \mathbf{X} \times \mathbb{R}$, and any permutation π of $\{1, \dots, n\}$,

$$A(z_1, \dots, z_n, z) = A\left(z_{\pi(1)}, \dots, z_{\pi(n)}, z\right).$$

The *conformal transducer* determined by a conformity measure A is defined as

$$Q(z_1, \dots, z_n, (x, y), \tau) := \frac{1}{n+1} \left|\{i = 1, \dots, n \mid \alpha_i^y < \alpha^y\}\right| + \frac{\tau}{n+1} \left|\{i = 1, \dots, n \mid \alpha_i^y = \alpha^y\}\right| + \frac{\tau}{n+1}, \quad (2)$$

where $(z_1, \dots, z_n) \in (\mathbf{X} \times \mathbb{R})^n$ is a training sequence, $x \in \mathbf{X}$ is a test object, and for each $y \in \mathbb{R}$ the corresponding *conformity scores* α_i^y and α^y are defined by

$$\alpha_i^y := A(z_1, \dots, z_{i-1}, (x, y), z_{i+1}, \dots, z_n, z_i), \qquad i = 1, \dots, n,$$
$$\alpha^y := A(z_1, \dots, z_n, (x, y)).$$

A function is a *conformal transducer* if it is the conformal transducer determined by some conformity measure. A *conformal predictive system* is a function which is both a conformal transducer and a randomized predictive system. We will also use the same terminology (randomized and conformal predictive systems, etc.) in the situation where n ranges over the natural numbers $\mathbb{N}$; e.g., a family $(Q_n)_{n \in \mathbb{N}}$ of randomized predictive systems $Q_n : (\mathbf{X} \times \mathbb{R})^{n+1} \times [0, 1] \to [0, 1]$ will also be referred to as a randomized predictive system.

A *conformal predictive distribution* (CPD) is (2) considered as a function of y, for a conformal predictive system Q. Property R1a ensures that this function is increasing, which is one of the defining properties of distribution functions. Property R1b says that it changes essentially from 0 to 1 as y increases from $-\infty$ to ∞, which is another defining property. Property R2 is the main property of validity for conformal predictive systems, which is satisfied automatically (Vovk et al., 2005). The last argument τ in $Q(z_1, \dots, z_n, (x, y), \tau)$ does not affect the value of Q much but still makes the property of validity R2 possible. In the terminology of Vovk et al. (2017), the *thickness* of (2) is typically $1/(n+1)$, meaning that (2) does not change by more than $1/(n+1)$ when τ ranges over $[0, 1]$ unless y is one of a finite number of points.

The following lemma gives a simple property of CPDs allowing us to integrate over them efficiently.

Lemma 1. *Any CPD*

$$Q^*(y) := Q(z_1, \dots, z_n, (x, y), \tau) \quad (3)$$

is a piecewise constant function that has at most $2n$ *points of discontinuity.*

Proof. Since $Q(z_1, \dots, z_n, (x, y), \tau)$ is a convex mixture of $Q(z_1, \dots, z_n, (x, y), 0)$ and $Q(z_1, \dots, z_n, (x, y), 1)$, namely,

$$Q(z_1, \dots, z_n, (x, y), \tau) = (1 - \tau) Q(z_1, \dots, z_n, (x, y), 0) + \tau Q(z_1, \dots, z_n, (x, y), 1),$$

it suffices to prove that $Q(z_1,\dots,z_n,(x,\cdot),0)$ and $Q(z_1,\dots,z_n,(x,\cdot),1)$ are piecewise constant functions that have at most n points of discontinuity. Let us consider, for concreteness, $Q(z_1,\dots,z_n,(x,\cdot),0)$. By definition (cf. (2)), this function is monotonically increasing and takes values in the set $\{0,1/(n+1),\dots,n/(n+1)\}$; therefore, it is piecewise constant and has at most n jumps. □

For any function $f:\mathbb{R}\to\mathbb{R}$, we define the integral

$$\int f(y)Q(z_1,\dots,z_n,(x,\mathrm{d}y),\tau)=\int f\mathrm{d}Q^*:=\sum_{y:\Delta Q^*(y)\neq 0} f(y)\Delta Q^*(y), \tag{4}$$

where $\Delta Q^*(y):=Q^*(y+)-Q^*(y-)$ is the jump of the CPD Q^*, as defined by (3), at y; by Lemma 1, the sum in (4) is finite.

3 The standard problem of decision making

In the rest of this note, with each label $y\in\mathbf{Y}$ and each decision $d\in\mathbf{D}$ we associate a von Neumann–Morgenstern utility $U(y,d)$. Formally, we are given a *utility function* $U:\mathbf{Y}\times\mathbf{D}\to\mathbb{R}$, which will always be assumed measurable. We will assume that the decision space is finite, $|\mathbf{D}|<\infty$.

Remark. This note's definitions and results can be extended to the case where U depends, additionally, on the object x, so that U becomes a function of three variables, $x\in\mathbf{X}$, $y\in\mathbf{Y}$, and $d\in\mathbf{D}$. However, this would require an extension of a result (Vovk, 2017, Theorem 3) that we will need in the proof of Theorem 2 below. Moreover, in applications, $U=U(y,d)$ does not usually depend on the object x.

We will need a stronger version of condition R1b: namely, in addition to (1), we will assume

$$\begin{aligned}\lim_{y\to-\infty}Q(z_1,\dots,z_n,(x,y),1)&=\frac{1}{n+1},\\ \lim_{y\to\infty}Q(z_1,\dots,z_n,(x,y),0)&=\frac{n}{n+1}.\end{aligned} \tag{5}$$

This condition is usually satisfied: consider, e.g., the standard definition

$$A(z_1,\dots,z_n,(x,y)):=y-\hat{y}, \tag{6}$$

where $\hat{y}$ is the prediction for y computed from $z_1,\dots,z_n$ and (x,y) by a method that is invariant with respect to permutations of $z_1,\dots,z_n$, or its variation

$$A(z_1,\dots,z_n,(x,y)):=\frac{y-\hat{y}}{\hat{\sigma}_y},$$

where $\hat{\sigma}_y>0$ is an estimate of the variability or difficulty of y computed from $z_1,\dots,z_n$ and (x,y) by a method that is invariant with respect to permutations

Algorithm 1 Conformal predictive decision making

Require: A training sequence $(x_i, y_i) \in \mathbf{Z}$, $i = 1, \dots, n$.
Require: A test object $x \in \mathbf{X}$.
1: **for** $d \in \mathbf{D}$ **do**
2: Create a new training sequence $(x_i, U(y_i, d))$, $i = 1, \dots, n$.
3: Find the CPD Q_d^* by (8) from this training sequence.
4: Compute the utility of d as $\int u Q_d^*(\mathrm{d}u)$.
5: **end for**
6: Return a $d \in \mathbf{D}$ with the largest utility.

of $z_1, \dots, z_n$. (Both methods may ignore y, of course.) Condition (5) ensures that different integrals (4) are comparable; even though they are not expected values (the total mass of Q^* is less than 1), they are nearly expected values (the total mass of Q^* is $n/(n+1)$) with the same shortfall of the total mass.

Informally, our task is, given a training sequence $z_1, \dots, z_n$, where $z_i = (x_i, y_i)$, and a test object $x \in \mathbf{X}$, to choose a suitable decision $d \in \mathbf{D}$ in view of the utility function U; the problem is that the utility $U(y, d)$ depends on the unknown label y of x. A *predictive decision-making system* (PDMS) is a measurable function $F : \mathbf{Z}^n \times \mathbf{X} \to \mathbf{D}$, or a family of such functions for all $n = 1, 2, \dots$; the intuition is that $F(z_1, \dots, z_n, x)$ is the decision recommended for the new object x based on the training set $z_1, \dots, z_n$. It is *randomized* when it depends, additionally, on a random number $\tau \in [0, 1]$ (representing internal coin tossing). The *regret* of a PDMS F (perhaps randomized) on an object x and training set $z_1, \dots, z_n$ under a probability measure P on $\mathbf{Z}$ is defined to be

$$R_F(z_1, \dots, z_n, x) := \max_{d \in \mathbf{D}} \int U(y, d) P(\mathrm{d}y \mid x) - \int U(y, F(z_1, \dots, z_n, x)) P(\mathrm{d}y \mid x), \tag{7}$$

where $P(\mathrm{d}y \mid x)$ is a regular conditional distribution (assumed to exist) for y given x under P. The F in the last integral in (7) may depend on the internal coin tossing (i.e., on $\tau \sim U_{0,1}$), in which case we write $R_F(z_1, \dots, z_n, x, \tau)$ for the left-hand side of (7). We are interested in PDMSs with small regret.

Our randomized algorithm is given as Algorithm 1. The CPD Q_d^* in line 3 is defined as

$$Q_d^*(u) := Q\big((x_1, U(y_1, d)), \dots, (x_n, U(y_n, d)), (x, u), \tau\big), \tag{8}$$

where Q is defined by (2) and assumed to satisfy (5).

In statistical decision theory, it is customary to use the *loss function* $L(y, d) := -U(y, d)$. (According to Berger 1993, "Statisticians seem to be pessimistic creatures who think in terms of losses.") Algorithm 1 in terms of losses is obtained by replacing U with L, replacing "utility" with "loss", and replacing "largest" with "smallest".

3.1 Basic classification

A basic special case is where $\mathbf{Y}$ is a finite set, $\mathbf{D} = \mathbf{Y}$, and the loss function is

$$L(y, d) := \begin{cases} 0 & \text{if } y = d \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

In this and next subsections we will be applying Algorithm 1 in terms of losses.

Let us check that, in the case of basic classification, Algorithm 1 becomes a version of the standard one-against-the-rest procedure (Vovk et al., 2005). For a fixed $d \in \mathbf{D}$, we create a new training set replacing label d by 0 and replacing all other labels by 1. The resulting CPD Q_d^* is not necessarily concentrated, or nearly concentrated, at one point (intuitively, such a point would represent the probability that the test label is different from d); however, we can still interpret $\int u Q_d^*(\mathrm{d}u)$ as the probability that the test label will be different from d. Now we predict the label d with the highest probability (by choosing the smallest $\int u Q_d^*(\mathrm{d}u)$).

3.2 Asymmetric classification

There are many cases where, unlike (9), different types of classification errors lead to different losses (for example, classifying an ill person as healthy is usually regarded a graver mistake than classifying a healthy person as ill). This corresponds to replacing 1 in (9) (which can now be interpreted as a square matrix of size $|\mathbf{Y}| \times |\mathbf{Y}|$) by different positive numbers.

3.3 Matrix games against nature

More generally, we can consider arbitrary finite sets $\mathbf{Y}$ and $\mathbf{D}$; then the utility function U (or loss function L) can be identified with a $|\mathbf{Y}| \times |\mathbf{D}|$ matrix.

4 An example application

To illustrate the application of Algorithm 1 in practice we use the Mushroom data set from the UCI repository (Dheeru and Karra Taniskidou, 2017). This data set contains observations on whether a mushroom is edible or not based on 22 observed attributes. We use an asymmetric utility function which penalizes eating a poisonous mushroom severely, as identified by the following matrix of utilities:

	eat	don't eat
edible	1	0
not edible	−10	1

We use the standard conformity measure (6), where $\hat{y}$ is the 1 Nearest Neighbour prediction for the label of x based on the Hamming distance. In Figure 1 we contrast the observed mean utility versus training set size for Algorithm 1 and

simply eating only mushrooms predicted to be edible. We draw 1000 random balanced training sets for each of a range of sizes ($\{4, 6, \dots, 20\}$), evaluating the performance of the two procedures on each training set using random balanced test sets of size 10.

One clearly observes that Algorithm 1 appropriately offsets the decision not to eat potentially poisonous mushrooms but that this benefit reduces with increasing training size as predictivity improves overall.

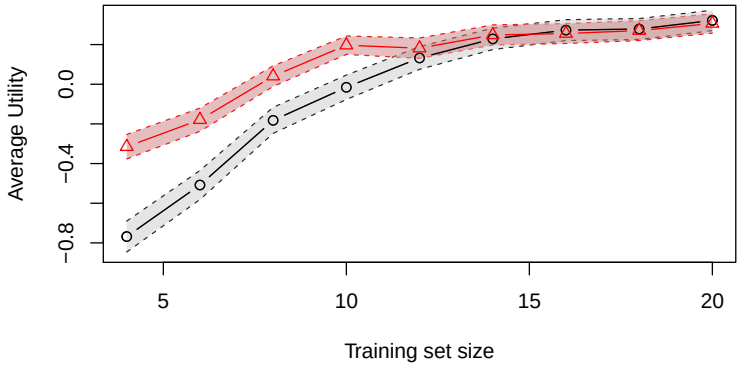

Figure 1: Average utility vs training set size for Algorithm 1 ($\triangle$) and simply eating only mushrooms predicted to be edible ($\circ$). Shaded areas indicate 95% confidence intervals on the mean.

5 Asymptotically efficient decision making

In this section we state a simple corollary of a result in Vovk (2017) showing that asymptotically we can choose the best possible decisions under weak assumptions about the object space $\mathbf{X}$ and the label space $\mathbf{Y}$.

Let us say that a randomized predictive system Q is *consistent* for a probability measure P on $\mathbf{X} \times \mathbb{R}$ if, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\int f \mathrm{d}Q_n - \int f(y) P(\mathrm{d}y \mid x_{n+1}) \to 0 \qquad (n \to \infty) \tag{10}$$

in probability, where:

- Q_n is the predictive distribution $Q_n : y \mapsto Q(z_1, \dots, z_n, (x_{n+1}, y), \tau_n)$ (for a given τ_n) output by Q as its forecast for the label y_{n+1} of x_{n+1} based on the training set $(z_1, \dots, z_n)$, where $z_i = (x_i, y_i) \in \mathbf{X} \times \mathbb{R}$;
- $P(\mathrm{d}y \mid x_{n+1})$ is a regular conditional distribution of y given $x = x_{n+1}$ under $(x, y) \sim P$;
- $z_n \sim P$ and $\tau_n \sim U_{0,1}$, $n = 1, 2, \dots$, are assumed independent.

It is clear that this notion of consistency does not depend on the choice of the version of the regular conditional distribution $P(\mathrm{d}y \mid x)$ in (10). Notice that the observations $z_1, z_2, \ldots$ are assumed to be generated from P in the IID fashion, and the internal coin tosses $\tau_1, \tau_2, \ldots$ in Q are assumed to be independent of them. The randomized predictive system Q is *universally consistent* if it is consistent for any probability measure P on $\mathbf{X} \times \mathbb{R}$.

A randomized PDMS is *asymptotically efficient* if, for any probability measure P on $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, its regret $R_F(z_1, \ldots, z_n, x_{n+1}, \tau_n)$ on x_{n+1} and $z_1, \ldots, z_n$ under P tends to 0 in probability when $z_i = (x_i, y_i) \in \mathbf{Z}$, $i = 1, 2, \ldots$, are produced from P and the random numbers τ_i are produced from $U_{0,1}$, all independently.

Theorem 2. *Suppose, in addition to the decision space* $\mathbf{D}$ *being finite, that* $\mathbf{X}$ *is a standard Borel space and the utility function* $U : \mathbf{Y} \times \mathbf{D} \to \mathbb{R}$ *is bounded. Asymptotically efficient PDMSs exist; in particular, Algorithm 1 is an asymptotically efficient randomized PDMS whenever it is based on a universally consistent randomized predictive system (which exists under these assumptions).*

Proof. It suffices to prove that, for each $d \in \mathbf{D}$,

$$\int U(y, d) P(\mathrm{d}y \mid x_{n+1}) - \int u Q_d^*(\mathrm{d}u) \to 0 \qquad (n \to \infty) \tag{11}$$

in probability, where Q_d^* is computed as in Algorithm 1 for the training set $z_1, \ldots, z_n$, test object x_{n+1}, and random number τ_n; indeed, under (11) and by (7), the requirement

$$R_F(z_1, \ldots, z_n, x_{n+1}, \tau_n) \to 0$$

can be rewritten as

$$\max_{d \in \mathbf{D}} \int u Q_d^*(\mathrm{d}u) - \int u Q_{F(z_1, \ldots, z_n, x_{n+1}, \tau_n)}^*(\mathrm{d}u) \to 0,$$

which is true (even with "$\to$" replaced by "$=$") by definition for F computed by Algorithm 1. Fix any $d \in \mathbf{D}$ and let P' be the image of the probability distribution P under the mapping $(x, y) \in \mathbf{Z} \mapsto (x, U(y, d)) \in \mathbf{X} \times \mathbb{R}$. Then (11) can be rewritten as

$$\int u P'(\mathrm{d}u \mid x_{n+1}) - \int u Q_d^*(\mathrm{d}u) \to 0 \qquad (n \to \infty),$$

and so the conclusion follows from the universal consistency of the randomized predictive system used in Algorithm 1, since the identity function $u \mapsto u$ is continuous and, under $P'(\mathrm{d}u \mid x_{n+1})$, bounded with probability one; the assumption that $\mathbf{X}$ is a standard Borel space is used in the statement of universal consistency (Vovk, 2017, Theorem 3). □

A universally consistent randomized predictive system is constructed in Vovk (2017) explicitly, and it is, of course, a conformal predictive system. Theorem 2 does not use its property of validity (R2) but this property is likely to improve the quality of predictions as measured by the utility function U.

6 Connections with existing literature and dangers of overfitting

In the first two subsections of this section we will briefly discuss classical statistical decision theory and its predictive counterpart, and the classical theory of confidence distributions and its predictive counterpart. Both classical theories are concerned with the value of a statistical parameter and, therefore, are rarely useful in the nonparametric framework of mainstream machine learning (since the parameter space, the set of all probability measures on the observation space $\mathbf{Z}$, then becomes too large). Their predictive versions replace the parameter by a future random variable. In the final two subsections we briefly compare conformal predictive decision theory with statistical learning theory and discuss the dangers of overfitting in both theories.

6.1 Statistical decision theory

Statistical decision theory was introduced by Wald (1950), who was in part inspired by developments in game theory (von Neumann and Morgenstern, 1953). The latter book introduced utilities (von Neumann and Morgenstern, 1953, Appendix, added in the second edition). The main difference between game-theoretic decision theory and statistical decision theory is that in the former the opponent is intelligent whereas in the latter it is of a neutral nature. A minor difference is that the former prefers the language of utilities, and the latter that of losses (which can be defined as minus utilities; we will ignore the difference in this appendix). The existing literature on statistical decision theory is massive, often combines ideas of decision theory with Bayesian ideas, and contains such well-known books as Berger (1993), Schervish (1995, Chapter 3), and Bernardo and Smith (2000). In the classical version the utility function $U = U(\theta, d)$ depends on a statistical parameter θ and decision d (see, e.g., Berger 1993, Section 1.2). In the predictive version (see, e.g., Berger 1993, Section 2.4.4) the utility function $U = U(y, d)$ depends on a future random variable y and decision d. It is often argued (in, e.g., Berger 1993, Section 2.4.4) that the predictive version can be reduced to the classical version, but the usefulness of such a reduction is limited in the nonparametric context of this note.

6.2 Confidence and predictive distributions

As we mentioned in Section 1, predictive distributions were introduced by Schweder and Hjort (2016, Chapter 12) and Shen et al. (2018). Both groups of authors were motivated by confidence distributions for a parameter value; moreover, whereas the latter group used the term "predictive distributions", the former used the longer "prediction confidence distributions". The term "confidence distribution" was introduced by Cox (1958), but the notions of both confidence and predictive distributions had been widely used by Fisher under the rubric of "fiducial distributions". The formal definition of confidence distributions is due to Schweder and Hjort (2002, Definition 1) and Singh et al.

(2005, Definition 1.1) (see also the influential review by Xie and Singh 2013); the key element of the definition is property R2 (see Section 2 above). In the nonparametric context of this note we only use predictive distributions.

6.3 Statistical learning theory

The standard problem of supervised learning (see, e.g., Vapnik 2000, Section 1.2) can be stated in terms of decision making. We still have an object space $\mathbf{X}$, a label space $\mathbf{Y}$, a decision space $\mathbf{D}$, and a loss function $L : \mathbf{Y} \times \mathbf{D} \to \mathbb{R}$. We are also given a *learning machine* $F : \mathbf{X} \times \Lambda \to \mathbf{D}$, where Λ is a parameter space; for each parameter $\alpha \in \Lambda$, the function $F(x, \alpha)$ of $x \in \mathbf{X}$ can be considered as a decision rule giving a recommended decision $d := F(x, \alpha)$ in view of the observed object $x \in \mathbf{X}$. The goal is to minimize the *risk functional*

$$R(\alpha) := \int L(y, F(x, \alpha)) P(\mathrm{d}x, \mathrm{d}y),$$

where P is the unknown data-generating distribution. One way of solving this problem is the *empirical risk minimization principle*, which recommends minimizing $R(\alpha)$ after replacing P by the empirical probability distribution. One way of controlling overfitting uses the notion of VC dimension (overfitting is a much more serious problem in the case of statistical learning theory: e.g., the assumption of a finite Λ would be extremely restrictive, whereas our assumption of a finite $\mathbf{D}$ still allows, e.g., any classification problems, and the example given in the next subsection is rather exotic).

Conformal predictive decision making is different from statistical learning theory in that it does not need a learning machine and that it is based on predictive distributions that satisfy a small-sample property of validity (namely, R2). In this note we do not state directly any small-sample properties of validity of conformal predictive decision making and only state a result about its asymptotic efficiency.

6.4 Dangers of overfitting in our theory

In the main part of this note we only discussed the case of a finite decision space $\mathbf{D}$. One problem for an infinite, or finite but very large, $\mathbf{D}$ is the possibility of "overfitting", where one of the decisions $d \in \mathbf{D}$ can lead to a small loss simply by chance. As a simple example, consider a probability measure P on an observation space $\mathbf{Z} = \mathbf{X} \times \mathbf{Y} = \mathbf{X} \times \mathbb{R}$ such that the conditional distribution of the label $y \in \mathbf{Y}$ given the object $x \in \mathbf{X}$ is always continuous (for some version of the conditional distribution). The decision space is infinite, namely the set of all finite subsets of the label space $\mathbf{Y} = \mathbb{R}$. The utility function is

$$U(y, d) := \begin{cases} 1 & \text{if } d = \emptyset \\ 2 & \text{if } y \in d \\ 0 & \text{otherwise.} \end{cases}$$

Given a training sequence $(x_1, y_1), \ldots, (x_n, y_n)$, Algorithm 1 will return (for a reasonable conformity measure) a decision $d \supseteq \{y_1, \ldots, y_n\}$, which is clearly suboptimal; the optimal decision would be $d := \emptyset$.

The danger of overfitting increases when the underlying algorithm used in the method of conformal prediction is randomized and is very substantially affected by randomness (the random choice of τ in (2) typically does not affect the chosen decision and so does not contribute to overfitting). This effect is akin to the possibility of violating the property of validity by aggregated conformal predictors, as reported in Linusson et al. (2017).

In principle, the methods of this note are applicable to complicated decision spaces $\mathbf{D}$, such as the set of all probability measures on a finite label space $\mathbf{Y}$ (the problem of probabilistic classification), or even the set of all probability measures on $\mathbf{Y} := \mathbb{R}^d$ for $d \in \{2, 3, \ldots\}$ (the problem of multi-dimensional probabilistic regression; the case $d = 1$ is covered already by the basic method of Section 2). Apart from dangers of overfitting, computational problems can be expected to be severe, and finding efficient implementations for specific underlying algorithms is an interesting direction of further research.

7 Conclusion

We extend the method of conformal prediction to make it applicable to decision making. Our hope is that this extension will prove to be useful in practice, and this is perhaps the most important direction of further research. We have limited ourselves to analysing a given batch of data, without attempting active learning, and this limitation has made it possible to develop a fairly systematic theory.

Acknowledgments

Thanks to Ola Engkvist, Lars Carlsson, Alex Gammerman, and Valery Manokhin for useful discussions. We are grateful to the three anonymous reviewers of the conference version of this paper for helpful comments. This work has been supported by the EU Horizon 2020 Research and Innovation programme (grant 671555).

References

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1993.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 2000.

David R. Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

Tore Schweder and Nils L. Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29:309–332, 2002.

Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge, UK, 2016.

Jieli Shen, Regina Liu, and Minge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140, 2018.

Kesar Singh, Minge Xie, and William E. Strawderman. Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33:159–183, 2005.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, third edition, 1953. First edition: 1944.

Vladimir Vovk. Universally consistent predictive distributions, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 18, September 2017.

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

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Minge Xie. Nonparametric predictive distributions based on conformal prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 17, April 2017.

Abraham Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

Minge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81: 3–39, 2013.