Reducing the Error Rate by Refusing to Guess: a Conjugate Conformal Predictor Approach

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

This works shows both theoretically and empirically how to reduce the error rate of state-of-the-art machine learning algorithms by refusing to make predictions in certain cases even when the underlying algorithms do. Intuitively, our new Conjugate Prediction approach estimates the likelihood that a prediction will be in error and when that likelihood is high, the approach refuses to make a prediction. Unlike other approaches, we can probabilistically guarantee an error rate on predictions we do make (called decisive predictions). Empirically on seven diverse data sets (chosen for their size), our method can probabilistically guarantee to reduce the error rate to 1/4 of what it is in the state-of-the-art machine learning algorithm at a cost of under 20% refusals. In practice, the error rate is even lower than the guarantee.

1 INTRODUCTION

Motivation Applications Organization & Contributions

2 PROBLEM DESCRIPTION & BACKGROUND

In the following subsection, we present the basic data assumptions and the problem of classification with refuse option. We also briefly discuss some uses of the refuse option from the literature. Next, in Subsection 2.2., we focus on conformal prediction framework. We provide a description of the algorithm with some theoretical guarantees that we will prove their counterparts in Section 3.

2.1 Classification with Refuse Option

In this paper, we assume the classical i.i.d. (independent and identically distributed) data model: i.e., we are given $m+1$ independent data points from an unknown but fixed distribution $\mathcal{D}$, where each point Z_i consists of an object-label pair $Z_i = (X_i, Y_i)$ and where the labels come from a finite set $\mathcal{Y}$.

We provide the first m points $Z_1, \ldots, Z_m$ and the $m+1^{st}$ object X_{m+1} to a classifier and request a prediction for the label Y_{m+1}. The classifier either makes a prediction $\hat{Y}_{m+1}$ or refuses to make one. A refusal to predict corresponds in our notation to $\hat{Y}_{m+1} = \varnothing$ and we say a prediction is *decisive* if it is not refused. A decisive prediction is made but is incorrect, then that corresponds to $\hat{Y}_{m+1} \notin \{Y_{m+1}, \varnothing\}$.

Why are we introducing refuse option! Give some papers and contexts then refer to the next subsection for the motivation of controlling the error rate

2.2 Controlling the Probability of Error

combine with the previous section

Our goal is to control the error probability when a classifier makes a prediction. Given a data sequence, we train a classifier to predict the labels of unseen data coming from the same distribution. While doing that, we try to gain an accurate estimate of the error probability. Given those estimates, we can bound the error probability on predicted data points to any prespecified value we desire.[1] Our results and algorithms

Preliminary work. Under review by AISTATS 2016. Do not distribute.

[1] The emphasis in the literature is on bounding the error probability in terms of the number of errors the classifier makes on the training set -via generalization bounds- or on a hold-out set. An accessible and instructive literature survey of such research is (Langford, 2005).

are based on the "conformal prediction framework", explained in more detail in the following subsection.

For the sake of brevity, in the rest of the paper we drop the indices of $Z_{m+1} = (X_{m+1}, Y_{m+1})$ and $\hat{Y}_{m+1}$. Also, we denote the training sequence $Z_1, \ldots, Z_m$ with Z_1^m, while the collection of these training points are denoted by Σ_m. Note the latter is a multiset, i.e., several components of the collection may be identical.

2.3 Conformal Prediction

Conformal predictors were introduced by Vovk and Gammerman in ... Conformal predictors take a pre-specified error rate ϵ as input and generate a subset of potential labels, such that the true label falls into this set with probability more than $1 - \epsilon$.

In this subsection, we describe the mechanics of the inductive conformal predictors and present the error guarantees provided in the literature. For a detailed derivation, various extensions and practical applications of conformal predictors we refer the reader to (Vovk et. al, 2005), (...), and (....).

Intuitively, each conformal predictor is characterized by a non-conformity score, a function that assesses the distance between a given point (Z) and a set of points (Σ). In the rest of the paper we denote this scoring functions with $A(Z, \Sigma) \in \mathbb{R}$ which increases as Z differs more from Σ. Typically, non-conformity scores are built upon prediction algorithms such as:

- KNN: The distance between Z and the closest point in Σ with label Y.
- SVM: The negative distance between Z and the decision boundary trained on Σ.
- Kernel estimators, Logistic regression, ... (probability estimators): One minus the probability estimate for Z based on Σ.
- Random Forest: Fraction of the trees that misclassifies Z in a forest trained on Σ.

For further examples of non-conformity scores we refer the reader to Chapters 3 and 4 of (Vovk et. al, 2005).

An inductive predictor first splits the training set Σ_m into two partitions i) a core training set Σ_{trn} and ii) a calibration set Σ_{cal}. Then, it computes the non-conformity scores relative to the core training set for each point in the calibration set:

$$\mathcal{A} = \{\alpha \ : \ \alpha = A(z, \Sigma_{trn}), \ \forall z \in \Sigma_{cal}\}.$$

Finally, for each potential label $y \in \mathcal{Y}$, it calculates the score $\alpha^y = A((X, y), \Sigma_{trn})$ and includes this label into the prediction set if this score is less than a fraction ϵ of the values in $\mathcal{A}$. We can see this step as setting a threshold α^* such that

$$\frac{|\{\alpha : \alpha \in \mathcal{A}, \ \alpha \leq \alpha^*\}|}{|\mathcal{A}| + 1} \geq 1 - \epsilon,$$

and adding y to the prediction set if and only if $\alpha^y \leq \alpha^*$. Intuitively, if multiple labels enter into the prediction set then we interpret that as a refusal. In the case none of the labels are included to the prediction set, we instead declare an error. This type of errors implies that the error rate is chosen too high or indicate an outlier. In Section 3.1., we introduce conjugate scores as a mean to generate predictions on such cases as well.

Below, we present the conditional and unconditional coverage guarantees for the inductive conformal predictors. As discussed in the previous subsection, only the conditional coverage operationally meaningful in the off-line scenario and in the next section we will develop algorithms to control error probabilities instead of coverage probabilities.

Theorem 2.1. *(Vovk, 2012) For any target error rate $0 < \epsilon < 1$ and any inductive conformal predictor described above, by denoting the prediction set as $\Gamma \subseteq \mathcal{Y}$:*

1. *The probability that the true label is not covered is bounded by ϵ,*

$$P(Y \notin \Gamma) \leq \epsilon.$$

2. *For any confidence parameter $0 < \delta < 1$, the following inequality holds with probability more than $1 - \delta$:*

$$P(Y \notin \Gamma \mid Z_1^m) \leq \epsilon + \sqrt{\frac{-\log \delta}{2|\Sigma_{cal}|}}.$$

Comparison, inductivist objection

Remark Note that, in (Vovk, 2012) a tigther bound for the second part of the theorem is also presented, but (as also noted there) this looser bound implies a simple modification to the original algorithm. To guarantee a coverage probability smaller than ϵ with confidence $1 - \delta$, all we need to do is decreasing the target error rate to $\epsilon - \sqrt{\frac{-\log \delta}{2|\Sigma_{cal}|}}$.

3 INDUCTIVE CONJUGATE PREDICTION

In this section, we introduce a variation of conformal predictors to control the error probability for the non-refused predictions (hereafter called *decisive predictions*: *Anil: please decide on one of these and write just that. It's ok to depend on the training set.* $P\left(\hat{Y} \neq Y \mid \hat{Y} \neq \varnothing\right)$ or $P\left(\hat{Y} \neq Y \mid \hat{Y} \neq \varnothing,\ Z_1^m\right)$ instead of $P\left(\hat{Y} \neq Y\right)$ or $P\left(\hat{Y} \neq Y \mid Z_1^m\right)$.

Discuss why these are more relevant!

We first introduce conjugate scores to make sure our algorithm doesn't make empty predictions, next in Subsection 3.2, we introduce the naive inductive conjugate predictors as a tool to control the unconditional error probability. Finally, in Section 3.3, we introduce our algorithm that controls the training conditional error probability.

3.1 Conjugate Scores

In the classical conformal predictor framework, once the predictor is trained and calibrated on Z_1^m, it first calculates a non-conformity score for a new object X by assuming the missing label is y: $\alpha^y = A\left((X, y), \Sigma\right)$. Next, it compares α^y with a threshold α^* (computed from the calibration set such that only a fraction ϵ of the calibration points gives errors) to decide if the label y should be included into the prediction set or not. After this process is repeated for each potential label in $\mathcal{Y}$, the final prediction set is returned. However, this approach sometimes leads to empty prediction sets, especially when the target error rate is too high.

As a simple remedy for this problem, for a given non-conformity score $\alpha^y = A\left((X, y), \Sigma\right)$, we introduce a corresponding *conjugate score* as:

$$\beta^y = \max_{y' \neq y} A\left((X, y'), \Sigma\right)$$

and add y to the prediction set only if the corresponding α^y is less than both the threshold α^* and β^y.

This "conjugate" modification of conformal prediction ensures that the most conforming label is always included in the prediction set while decreasing the number of refused points because now we can choose a smaller α^* and still get no more errors in the calibration set(see Figure 1). A detailed study of conjugate scores is given in (Kocak et. al, 2016).

3.2 Controlling Unconditional Error Probability

In this section, we present the inductive conjugate prediction (from now on abbreviated as *ICP*) algorithm and give an bound on the probability of error on the decisive predictions. The high-level pseudocode of the ICP is given in Algorithm 1 below, and the details are presented in Algorithms 2 and 3.

This figure intentionally left non-blank

Figure 1: comparison of conformal and conjugate predictors, make sure this figure clarifies the function of α^*!

As the first step, ICP splits the training data into two non-overlapping sets as in the case of conformal predictors: the core training set Σ_{trn} and the calibration set Σ_{cal}. Next, the algorithm uses a conjugate predictor trained on the core training set with a fixed threshold α^* (see Algorithm 2 and Figure 1). In particular, in step 2, we choose the minimum α^* that leads to an empirical error rate less than ϵ on the calibration set – explained in detail below –, and in step 3, we employ this threshold to make the actual prediction $\hat{Y}$.

Calibration: In the calibration step, we start by setting the threshold to minimum possible value (either $-\infty$ or 0 depeding on the non-conformity score), and compute the fraction of errors among the decisive predictions on the calibration set, as an hold-out estimate of the error probability on the new data. Next, we gradually increase the threshold α^* till the computed estimate fall below the target error rate, see Eq. (1).

Dennis: The reason for the conservative estimate in Eq. 1 will become apparent in the proof of Thm. 3.1, but I am not sure if we should mention it here?

Algorithm 1 Inductive Conjugate Prediction

Input: training data Σ_m, target rate ϵ, test object X

Output: predicted label $\hat{Y}$ ($\varnothing$ stands for refusal)

```
1: Split the training data into two:
   Core training data: $\Sigma_{trn} \leftarrow \{Z_1, \ldots, Z_l\}$
   Calibration data: $\Sigma_{cal} \leftarrow \{Z_{l+1}, \ldots, Z_m\}$
2: Calibrate: Choose minimum $\alpha^*$ such that
   $(E, C) = ICPScore\left(\Sigma_{trn}, \Sigma_{cal}, \alpha^*\right)$ satisfies
       $$\frac{E+1}{C+1} \leq \epsilon \qquad (1)$$
3: Predict: $\hat{Y} \leftarrow ICPPredict\left(\Sigma_{trn}, \alpha^*, X\right)$
```

Next, we prove that the probability of error for an ICP on decisive predictions is upper-bounded by the targer error rate ϵ. Note that this probability is calculated over the training set as well. *Tentative:* the next subsection, we will prove the conditional counterpart of this bound and modify ICP algorithm to bound the training set conditional probability in a natural way.

Algorithm 2 $\hat{Y} = ICPPredict\left(\Sigma_{trn}, \alpha^*, X\right)$
Input: training set Σ_{trn}, threshold α^*, test object X
Output: predicted label $\hat{Y}$ ($\varnothing$ represents refusal)

```
$\Gamma \leftarrow \emptyset$
for $y \in \mathcal{Y}$ do
    $\alpha^y \leftarrow A\left(\Sigma_{trn}, (X, y)\right)$
    $\beta^y \leftarrow \min_{y' \neq Y_i} A\left(\Sigma_{trn}, (X_i, y')\right)$
    if $\alpha^y \leq \max\left(\beta^y, \alpha^*\right)$ then
        $\Gamma \leftarrow \Gamma \bigcup \{y\}$.
if $\Gamma$ is a singleton then
    $\hat{Y} \in \Gamma$
else
    $\hat{Y} = \varnothing$
```

Algorithm 3 $(E, C) = ICPScore\left(\Sigma_{trn}, \Sigma_{cal}, \alpha^*\right)$
Input: training set Σ_{trn}, calibration set Σ_{cal}, threshold α^*
Output: error count E, decisive prediction count C

```
$E, C \leftarrow 0, |\Sigma_{cal}|$
for each $Z \in \Sigma_{cal}$ do
    $\hat{Y} \leftarrow ICPPredict\left(\Sigma_{trn}, \alpha^*, X\right)$
    if $\hat{Y} = \varnothing$ then
        $C \leftarrow C - 1$
    else if $Y \neq \hat{Y}$ then
        $E \leftarrow E + 1$
```

Theorem 3.1. *If an inductive conjugate predictor given in Algorithm 1 is fed with (Σ_m, ϵ, X) and generates the output $\hat{Y}$, then the probability of error of a decisive prediction is less than the target error rate ϵ:*

$$P\left(\hat{Y} \neq Y | \hat{Y} \neq \varnothing\right) \quad \leq \quad \epsilon. \tag{2}$$

Proof. We first define the following:

- $\Sigma = \Sigma_{cal} \cup \{Z\}$: The union of calibration set and the test point.
- α_z: The threshold value we would get at the calibration step, if the calibration set were $\Sigma - \{z\}$.

Next, we can re-write the left-hand-side of Eq. (2) by using the total probability law, and the definition of conditional probability as follows:

$$\begin{aligned}
&= \mathbb{E}_{\Sigma}\left[\frac{P\left(Y \notin \{\hat{Y}, \varnothing\} |\ Z \in \Sigma\right)}{P\left(\hat{Y} \neq \varnothing |\ \alpha^* = \alpha_Z, Z \in \Sigma\right)}\right] \\
&= \mathbb{E}_{\Sigma}\left[\mathbb{E}_{\alpha|\Sigma}\left[\frac{P\left(Y \notin \{\hat{Y}, \varnothing\} |\ \alpha^* = \alpha, Z \in \Sigma\right)}{P\left(\hat{Y} \neq \varnothing |\ \alpha^* = \alpha, Z \in \Sigma\right)}\right]\right] \\
&\leq \mathbb{E}_{\Sigma}\left[\max_{z \in \Sigma} \frac{P\left(Y \notin \{\hat{Y}, \varnothing\} |\ \alpha^* = \alpha_z, Z \in \Sigma\right)}{P\left(\hat{Y} \neq \varnothing |\ \alpha^* = \alpha_z, Z \in \Sigma\right)}\right].
\end{aligned}$$

The second line follows from the tower law, and the third line follows by replacing the expectation with the maximum. In the following we denote this maximizing z value as z^*. Note that, the quantity in the expectation does not depend on the calibration set anymore since the threshold value α^* is only a function of Σ, instead of Σ_{cal}. Thus we can compute this term by exploiting the fact the points in Σ are equiprobable:

$$\ldots \quad = \quad \mathbb{E}_{\Sigma}\left[\frac{E'}{C'}\right] \tag{3}$$

where $(E', C') = ICPScore\left(\Sigma_{trn}, \Sigma, \alpha_{z^*}\right)$.

Finally, we compare (E', C') with

$$(E'', C'') = ICPScore\left(\Sigma_{trn}, \Sigma - \{z^*\}, \alpha_{z^*}\right).$$

Since the former has only one more data point in its calibration set, the error and decisive prediction counts can differ at most 1. Hence, we have

$$\frac{E'}{C'} = \left\{\frac{E''}{C''}, \frac{E''+1}{C''+1}, \frac{E''}{C''+1}\right\}.$$

Consequently, continuing from Eq (3)

$$\begin{aligned}
\ldots \quad &\leq \quad \mathbb{E}_{\Sigma}\left[\frac{E''+1}{C''+1}\right] \\
&\leq \quad \mathbb{E}_{\Sigma}\left[\epsilon\right] \quad = \quad \epsilon.
\end{aligned} \tag{4}$$

The last step follows from the definition of calibration step, and concludes the proof. □

maybe short remark

3.3 Controlling Training Conditional Error Probability

As noted in the discussion following Theorem 2.1., the bound given in Theorem 3.1. gives information about the error probability for the decisive predictions averaged over the training set. .... In this subsection instead we will focus on the training conditional error probability over the decisive predictions, i.e.,

$$P\left(\hat{Y} \neq Y \mid \hat{Y} \neq \varnothing,\ Z_1^m\right).$$

Before presenting our main result, Theorem 3.2., we start with some notation and observations.

More Notation & Observations:

i.

$$\alpha\left(\epsilon\right) = \inf\left\{\alpha : P\left(\hat{Y} \neq Y | \hat{Y} \neq \varnothing, \alpha^* = \alpha\right) \leq \epsilon\right\}$$

ii.

$$e\left(\epsilon\right) = P\left(\hat{Y} \notin \{Y, \varnothing\} \mid \alpha^* = \alpha\left(\epsilon\right)\right)$$

$$c\left(\epsilon\right) = P\left(\hat{Y} \neq \varnothing \mid \alpha^* = \alpha\left(\epsilon\right)\right)$$

Theorem 3.2. *For any $\delta_1, \delta_2 \in (0, 1)$, the following inequality holds with probability more than $1 - \delta_1 - \delta_2$*

$$P\left(\hat{Y} \neq Y \mid \hat{Y} \neq \varnothing, Z_1^m\right) \quad \leq \quad \frac{e\left(\epsilon + \Delta_1\right)}{c\left(\epsilon - \Delta_2\right)}. \qquad (5)$$

where $\Delta_1 = O\left(\sqrt{\frac{\log n}{n}}\right)$ *and* $\Delta_2 = O\left(\sqrt{\frac{\log n}{n}}\right)$.

Proof. The proof simply follows combining the Propositions 3.2 and 3.3. with union bound and defining $\Delta_1 = \epsilon' - \epsilon$, $\Delta_2 = \epsilon - \epsilon''$, i.e.,

$$P\left(\alpha\left(\epsilon - \Delta_2\right) > \alpha^* > \alpha\left(\epsilon + \Delta_1\right)\right)$$

with probability more than $1 - \delta_1 - \delta_2$. Finally by noting that both e and c are monotone increasing functions, we conclude the proof by the following inequalities:

$$P\left(\hat{Y} \notin \{Y, \varnothing\} \mid Z_1^m\right) \leq e\left(\epsilon + \Delta_1\right),$$

$$P\left(\hat{Y} \neq \varnothing \mid Z_1^m\right) \geq c\left(\epsilon - \Delta_2\right).$$

□

Remark ....

Proposition 3.3. *For any $0 < \delta_1 < 1$ and ϵ' satisfying*

$$\epsilon' \geq \epsilon + O\left(\sqrt{\frac{\log n}{n}}\right),$$

following bound holds:

$$P\left(\alpha^* \leq \alpha\left(\epsilon'\right)\right) \leq \delta_1.$$

Proposition 3.4. *For any $0 < \delta_2 < 1$ and ϵ'' satisfying*

$$\epsilon'' \leq \epsilon - O\left(\sqrt{\frac{\log n}{n}}\right),$$

following bound holds:

$$P\left(\alpha^* \geq \alpha\left(\epsilon''\right)\right) \leq \delta_2.$$

4 NUMERICAL EXPERIMENTS

Table 1: Sample Table Title

Dataset Name	Number of Instances	Number of Features	Number of Classes
MNIST	a	a	10
Cover	a	a	7
Sensit	1	a	a
Connect-4	1	a	a
Letter	1	a	26
Cod-RNA	1	a	a
Sat-Image	1	a	a

5 DISCUSSIONS

This figure intentionally left non-blank

Figure 2: comparison of conformal and conjugate predictors, make sure this figure clarifies the function of α^*!

Table 2: $\delta = 0.01$, Entries in the table are in the form of percentages!

	Base Predictor	ϵ		$\epsilon/2$		$\epsilon/4$	
	Error Rate (ϵ)	Error	Refuse	Error	Refuse	Error	Refuse
MNIST	04.42 ± 0.15	04.45 ± 0.21	01.84 ± 0.39	02.24 ± 0.15	10.85 ± 0.48	01.12 ± 0.17	17.70 ± 0.67
Cover	08.54 ± 0.08	08.48 ± 0.11	04.63 ± 0.36	04.23 ± 0.11	20.94 ± 0.37	02.12 ± 0.06	35.55 ± 0.61
Sensit	16.65 ± 0.24	16.61 ± 0.34	01.14 ± 0.70	08.22 ± 0.33	31.85 ± 0.78	04.15 ± 0.34	57.13 ± 0.60
Connect-4	16.36 ± 0.22	15.34 ± 0.24	03.38 ± 0.93	08.06 ± 0.22	31.81 ± 0.84	04.06 ± 0.19	49.25 ± 0.78
Letter	004.48 ± 0.24	04.57 ± 0.37	08.12 ± 2.05	02.37 ± 0.24	21.22 ± 1.52	01.26 ± 0.09	29.13 ± 1.37
Cod-RNA	03.16 ± 0.06	03.17 ± 0.08	00.45 ± 0.09	01.58 ± 0.06	05.69 ± 0.15	00.80 ± 0.04	11.65 ± 0.20
Sat-Image	07.57 ± 0.48	07.70 ± 0.55	03.55 ± 1.07	03.86 ± 0.41	15.18 ± 1.07	01.96 ± 0.40	23.91 ± 2.17

6 FIRST LEVEL HEADINGS

6.1 Second Level Heading

6.1.1 Citations in Text

Citations within the text should include the author's last name and year, e.g., (Cheesman, 1985). References should follow any style that you are used to using, as long as their style is consistent throughout the paper. Be sure that the sentence reads correctly if the citation is deleted: e.g., instead of “As described by (Cheesman, 1985), we first frobulate the widgets,” write “As described by Cheesman (1985), we first frobulate the widgets.” Be sure to avoid accidentally disclosing author identities through citations.

One line space before the table title, one line space after the table title, and one line space after the table. The table title must be initial caps and each table numbered consecutively.

References

J. Alspector, B. Gupta, and R. B. Allen (1989). Performance of a stochastic learning microchip. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, 748-760. San Mateo, Calif.: Morgan Kaufmann.

F. Rosenblatt (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

G. Tesauro (1989). Neurogammon wins computer Olympiad. *Neural Computation* **1**(3):321-323.