Conjugate Conformal Prediction for Online Binary Classification

Mustafa A. Kocak
Tandon School of Engineering
New York University
Brooklyn, NY 11201
kocak@nyu.edu

Elza Erkip
Tandon School of Engineering
New York University
Brooklyn, NY 11201
elza@nyu.edu

Dennis E. Shasha
Courant Institute
New York University
New York, NY 10012
shasha@cs.nyu.edu

Abstract

Binary classification (rain or shine, disease or not, increase or decrease) is a fundamental problem in machine learning. We present an algorithm that can take any standard online binary classification algorithm and provably improve its performance under very weak assumptions, given the right to refuse to make predictions in certain cases. The extent of improvement will depend on the data size, stability of the algorithm, and room for improvement in the algorithms performance. Our experiments on standard machine learning data sets and standard algorithms (k-nearest neighbors and random forests) show the effectiveness of our approach, even beyond what is possible using previous work on conformal predictors upon which our approach is based. Though we focus on binary classification, our theory could be extended to multiway classification. Our code and data are available upon request.

1 INTRODUCTION

Reacting intelligently to incoming data lies at the heart of forecasting, trading, and many other applications. The simplest decision one has to make is binary (go/stop, buy/sell). However, in many cases one needs to assess the risk of a decision and refuse to make a decision at all if one is not confident enough. One of the well known methodologies for this task is using confidence predictors and declining to make a decision on ambiguous data points (Vovk, Gammerman, & Shafer, 2005). Intuitively, conformal predictors look at how previous predictions worked out for similar input data and let those results shape first whether to make a prediction at all and if so, which one to make. We extend existing conformal prediction approaches by permitting the use of multidimensional test statistics to provide more flexibility while keeping the theoretical guarantees of the original predictors. We apply this extended framework to the binary classification problem and show that our methods improve the performance of previous conformal predictors.

1.1 Related Work

Forecasting the upcoming data point in a data stream has been studied in econometrics, meteorology, finance, computer science, and many other disciplines (Box, Jenkins, Reinsel, & Ljung, 2015). In confidence predictors, the goal is to create a set of possible candidate outcomes such that the probability that the real outcome is not one of these candidates is less than a predetermined tolerance level. Some examples for the uses of confidence predictions include signal denoising (Ryabko & Ryabko, 2013), growth estimation for planning (Meade & Islam, 1995), among many similar forecast problems that requires bounds for the predicted values (Chatfield, 1993). Though most of the confidence prediction literature considers either parametric models or asymptotic results, Vovk, Gammerman and Shaffer (Vovk et al., 2005) introduced the conformal prediction framework, which provides exact finite-sample guarantees for exchangeable data sequences.

Binary classification problems where the classifier is allowed to decline making a decision has been studied both in online and offline settings. Optimal error-rejection tradeoffs are investigated under the names *selective classification* (El-Yaniv & Wiener, 2010, 2012), (Chaudhuri & Zhang, 2013) and *classification with reject option* (Denis & Hebiri, 2015), (Chow, 1970), (Herbei & Wegkamp, 2006).

Following this work, we focus on the conformal prediction approach to the online binary classification with reject option, and propose an extended framework to decrease the number of rejects while guaranteeing a small probability of error on the classified points.

1.2 Outline & Contributions

In Section 2, we describe our problem formally and provide some background information on online confidence prediction and conformal prediction. In Section 3, we extend the conformal prediction framework to multiple dimensions and then show how to preserve the theoretical error guarantees provided by the conformal framework, while rejecting less often.

The flexibility provided by this multidimensional framework is demonstrated in Section 4, where we introduce the notion of *conjugate conformal prediction*. Formally, a *conjugate conformal predictor* is a two dimensional conformal predictor derived from any given traditional conformal predictor. Intuitively, conjugate predictors not only compare the test point with the previous ones, but also compare the test point with its conjugate, i.e. the same point with its label flipped, to be able to make more aggressive predictions - not only when we have high confidence to accept the test point but also when we have high confidence to reject the conjugate point.

In Section 4.1, we define conjugate conformal predictors formally and prove their efficiency under mild stability assumptions on the set statistics used in the original predictor. Lastly, in Section 4.2. we present experimental results for classical conformal and conjugate conformal predictors on standard machine learning data sets from UCI ML Repository (Lichman, 2013) and standard machine learning algorithms (k nearest neighbors and random forests).

Finally, in Section 5 we conclude with a brief discussion of our results and planned future work.

2 PROBLEM SETUP & CONFORMAL PREDICTION

2.1 Data Model & Notation

In this work, we assume data points are revealed to the algorithm, one data point at a time. A data point generated at time t, consists of a feature vector x_t, which takes values from a feature space $\mathcal{X}$, and a label y_t, which takes values from a label space $\mathcal{Y}$. For the sake of brevity, we represent a data point with $z_t = (x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$. We refer to the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as the data space.

The only statistical assumption about the data generating process is the *exchangeability* of the data points, that is: for any positive integer N, the probability of observing a data sequence $z_1, \ldots, z_N$ is invariant under any permutation of the data points, π, i.e.

$$Pr\left(z_1, \ldots, z_N\right) \quad = \quad Pr\left(z_{\pi(1)}, \ldots, z_{\pi(N)}\right).$$

Many processes (or variants of these processes) satisfy this assumption. For example, stock price histories are not exchangeable, but stock price returns (percentage up or down over a given time period) are. Echangeability can be considered as a generalization of the more common i.i.d assumption. We refer the reader to (Schervish, 2012) and (Kallenberg, 2006) for a thorough discussion of this assumption.

In addition, we also make use of a source of randomness: uniform random variables $\tau_1, \tau_2, \ldots$ on the unit interval, which are independent of the data. This will be used to randomize the prediction algorithms in such a way as to achieve validity guarantees.

Lastly, we use multi-sets throughout the paper. In other words, each set may contain the same element multiple times, in particular, we denote the (multi-)set of the first t data points as $\sigma_t = \{z_1, \ldots, z_t\}$ and we don't require the data points to be distinct. In addition, we use the following variations of σ_t for the sake of brevity of exposition:

- $\sigma_t^{(i)} = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_t\}$ stands for the set of first t data points except the i^{th} one for any $i = 1, \ldots, t$.
- $\sigma_{t/y} = \{z_1, \ldots, z_{t-1}, (x_t, y)\}$ represents the set of first t data points assuming the t^{th} label is equal to y, for any $y \in \mathcal{Y}$.
- $\sigma_{t/y}^{(i)}$ is the set of data points in $\sigma_{t/y}$ with the exception of the i^{th} one, for $i = 1, \ldots, t$ and $y \in \mathcal{Y}$.

2.2 Online Prediction of Confidence

For the confidence prediction task, we assume the feature vector x_t is revealed at time t, but the corresponding label y_t is revealed only after the prediction is made and before the next feature vector x_{t+1} is revealed.

The task of the predictor is to predict a subset of the label space that contains the unseen label y_t with probability at least $1 - \epsilon$, for a given error tolerance level $\epsilon \in (0, 1)$. We denote the prediction set generated at time t as Γ_t^ϵ. Because Γ_t^ϵ is a set, we may be making a prediction of the form "the answer may be 1, 4, or 5". Formally:

Workflow for Online Confidence Prediction The fewer the errors, the more valid; the smaller the prediction set, the more efficient.

```
for $t = 1, 2, \ldots$ and given $\epsilon \in (0, 1)$ do
    Nature reveals $x_t$.
    Predictor calculates a prediction set $\Gamma_t^\epsilon \subseteq \mathcal{Y}$.
    Nature reveals $y_t$ and declares an error if $y_t \notin \Gamma_t^\epsilon$.
```

There are two main properties one expects from a good confidence predictor. The first is that the predictor should be *valid* (intuitively, the label value falls within the prediction set Γ_t^ϵ, fraction $1 - \epsilon$ of the time) and the other is that it is *efficient* (intuitively, $|\Gamma_t^\epsilon|$ is at least one but small).

The literature contains various ways of defining efficiency and validity measures, see (Vovk et al., 2014) for a detailed list. We use the following definitions:

- We call a confidence predictor *exactly valid* if the error events associated with its predictions occurs independently with probability ϵ. Additionally, we say a predictor is *conservatively valid* or simply *valid*, if it makes errors only on the data points on which some other exactly valid predictor makes errors. For further implications of this definition, see (Vovk et al., 2005).

- Since our results in Section 4 are focused on binary predictions, we use the cardinality of the predicted set, $|\Gamma_t^\epsilon|$, as a measure of efficiency. Particularly, we say the prediction at time t is *efficient* if $|\Gamma_t^\epsilon| = 1$. In binary classification setup, as in Section 4, an inefficient prediction at time t implies that the predictor chooses both possible values or none, effectively refusing to make a decision at t.

  Furthermore, we say a first predictor is more efficient than a second predictor at time t, if the first refuses to make a decision only if the second also refuses to make a decision at time t. Note that, such a comparison makes sense when both of the predictors are valid and have the same tolerance parameter, ϵ.

  For continuous label spaces, one can refer to the volume or size of the the prediction sets as a measure of efficiency (Lei, Robbins, & Wasserman, 2011).

2.3 Conformal Prediction

The main idea behind conformal predictors is to define a *nonconformity score* at time t between a candidate point $z = (x_t, y)$ for a candidate label $y \in \mathcal{Y}$ and the rest of the data $z_1, \ldots, z_{t-1}$, and to use it as a test statistic to decide whether a particular candidate label y will be included in the prediction set (candidate outcomes) or not. Intuitively, if data points similar to x_t have often mapped to y in the past, then y should belong to the prediction set for x_t.

The test statistic should take on smaller values the more probable the y is. One can create a non-conformity score based on a machine learning algorithm. Say that we employ the given algorithm, f, train it on the set σ_{t-1} and predict $\hat{y}_t = f_{\sigma_{t-1}}(x_t)$ as an estimate of y_t. Then we can derive a non-conformity score (also called the test statistic) to use at time t as $A_t(\sigma_{t-1}, z_t) = \phi\left(f_{\sigma_{t-1}}(x_t), y_t\right)$, where ϕ is any properly chosen loss function.

Because of the exchangeability assumption on the data, the conformity scores should be invariant to the order of data points in the training set σ_{t-1}. Therefore, the set of first t data points, σ_t, constitutes a complete sufficient statistic on the test statistic $A_t(\sigma_{t-1}, z_t)$. In our context, this means we can compute the exact probability distribution of the test statistic conditioned on σ_t. For a more detailed analysis of test statistics and how to compute their distribution under exchangeability, see Chapters 3.2 and 6.2 of (Cox & Hinkley, 1974).

Exploiting this idea leads to the following algorithm: calculate the non-conformity scores for each data point z_i that precedes t, based on the rest of the data preceding t plus the assumption that $y_t = y$ (the algorithm will do this one at a time for each label y for time t). We then calculate a confidence value, p_t^y, for the candidate label y as the fraction of the data points with greater non-conformity scores than the score of the last one (x_t, y). We will accept y into the confidence set provided this confidence value is greater than or equal to ϵ. Vovk (2005) has shown that, after some smoothing (using the additional source of randomness mentioned in section 2.1, the uniform random variable τ_t), this confidence value has a uniform distribution on $[0, 1]$ if the label y_t is really equal to y. The uniformity implies therefore that refusing to include y as a candidate point if its confidence value is less than ϵ would cause an error only with probability ϵ. The pseudocode is given in Algorithm 1, and further details are given in (Vovk et al., 2005).

Algorithm 1 (Smoothed) Conformal Prediction: The goal is to create a set of predicted values Γ_t^ϵ which covers the true label with probability $1 - \epsilon$, based on a confidence value p_t^y calculated from a given non-conformity measure A_t. In particular, p_t^y computes the fraction of data points with a larger non-conformity score than α_t^y, where τ_t is used to break the ties. The detailed definitions are in the text.

```
for $t = 1, 2, \ldots$ do
    $\Gamma_t^\epsilon \leftarrow \emptyset$
    for $y \in \mathcal{Y}$ do
        $y_t \leftarrow y$
        for $i = 1, \ldots, t$ do
            $\alpha_i^y \leftarrow A_t\left(\sigma_{t/y}^{(i)}, z_i\right)$
        $p_t^y = \left(|\{i : \alpha_i^y > \alpha_t^y, \}| + \tau_t |\{i : \alpha_i^y = \alpha_t^y\}|\right)/t$
        if $p_t^y \geq \epsilon$ then
            $\Gamma_t^\epsilon \leftarrow \Gamma_t^\epsilon \bigcup \{y\}$.
```

To appreciate the strength of the conformal predictors one can refer to the following results (Schafer & Vovk, 2008):

i. All (smoothed) conformal predictors are exactly valid (i.e. the error event at any time point is independent from others and occurs with probability ϵ) under the exchangeability assumption.

ii. If the data space is a Borel space, any given valid confidence predictor which is invariant to the order of the previous data points, there exists a conformal predictor that generates prediction sets not larger than the prediction sets generated by the given confidence predictor.

3 MULTIDIMENSIONAL CONFORMAL PREDICTION

In this section we will generalize the conformal prediction framework to multidimensional statistics and propose a principled extension to the algorithm presented above. This generalization is both simple and improves the efficiency of the formal prediction framework thus achieving practical improvements to standard algorithms.

The idea of the extension is to use non-conformity scores that take vector values instead of scalar ones, i.e. the range of A_t is $\mathbb{R}^d$ for some positive integer d. Such an approach may be helpful when we have several possible sets of data that may bear on a prediction. In such a scenario, one can use the non-conformity score of the candidate point to each of the several sets as components of a *non-conformity vector*.

The approach also helps in the application described in the next section, in which we focus on the case where the label space is binary, i.e. $\mathcal{Y} = \{0, 1\}$. In that setting, a one-dimensional non-conformity score $A_t\left(\sigma_{t-1}, (x_t, y)\right)$ and its conjugate $A_t\left(\sigma_{t-1}, (x_t, 1-y)\right)$ together provide a substantial improvement to the performance of the prediction compared to using just one score.

Just as in the one dimensional case, we assume that data points come from an exchangeable process and each component of the conformity vectors is invariant to the order of the points in the history, thus making the multi-dimensional conformity vectors exchangeable. Therefore, we can build a test statistic from them. However, in contrast to the scalar case, we don't have a linear order on these vectors, which complicates the decision of whether to include or exclude a label in the prediction set.

Instead of calculating a confidence value p^y for each $y \in \mathcal{Y}$ as before, we propose to select some subset of the d dimensional Euclidean space, which we call the *acceptance set* and denote it with S_t^y, for each y. We add the label y to the prediction set Γ_t^ϵ only if the corresponding nonconformity vector falls into the acceptance set, i.e. $y \in S_t^y$. Also, just as in the calculation of the one dimensional conformal prediction, we apply random smoothing on the boundary points of the acceptance set to guarantee exact validity. Specifically, we add y into the prediction set if the corresponding non-conformity vector is an interior point of the acceptance set, but if it is a boundary point of the set we include y iff τ_t is less than a specific value that is calibrated to the targeted error level. In Theorem 3.1 and the following Corollary 3.1.1, we provide some sufficient conditions on the acceptance sets to guarantee the validity of the associated predictor.

As an example, for the binary case, we propose to construct acceptance sets that include the points with smaller non-conformity scores than their conjugate scores in addition to some points with small non-conformity scores. This will satisfy the conditions given in Corollary 3.1.1 (See Figure 1). Such proposed acceptance sets will be investigated in detail in the next section.

The pseudocode for the described algorithm is given in Algorithm 2 below for generic acceptance sets. The following notation is used in the presentation of the algorithm to denote the acceptance sets:

- S_t^y: The acceptance set at time t for the prospective label y.
- $int\left(S_t^y\right)$: Interior points of the acceptance set.
- $\overline{int}\left(S_t^y\right)$: The set of points in the acceptance set, but not in the interior of it, i.e. $S_t^y / int\left(S_t^y\right)$.
- $\mathbf{v}_i^y$: The non-conformity vector for data point z_i, assuming $y_t = y$.
- Δ_t^y: Set of first t non-conformity vectors for $y_t = y$, i.e. $\{\mathbf{v}_1^y, \ldots, \mathbf{v}_t^y\}$.

Algorithm 2 Multidimensional Conformal Prediction: The goal is create a set of predicted outcomes Γ_t^ϵ based on d-dimensional statistics. See the definitions just above.

```
for $t = 1, 2, \ldots$ do
    $\Gamma_t^\epsilon \leftarrow \emptyset$
    for $y \in \mathcal{Y}$ do
        $y_t \leftarrow y$
        for $i = 1, \ldots, t$ do
            $\mathbf{v}_i^y \leftarrow A_t\left(\sigma_{t/y}^{(i)}, z_i\right)$
        Calculate $S_t^y \subseteq \mathbb{R}^d$ from $\sigma_{t/y}$
        if $\mathbf{v}_t^y \in int\left(S_t^y\right)$ then
            $\Gamma_t^\epsilon \leftarrow \Gamma_t^\epsilon \bigcup \{y\}$.
        if $\mathbf{v}_t^y \in \overline{int}\left(S_t^y\right)$ & $\tau_t \geq \frac{|S_t^y \cap \Delta_t^y| - (1-\epsilon)t}{|\overline{int}(S_t^y) \cap \Delta_t^y|}$ then
            $\Gamma_t^\epsilon \leftarrow \Gamma_t^\epsilon \bigcup \{y\}$.
```

The following theorem provides sufficient conditions for a sequence of acceptance sets using the above algorithm to guarantee that they lead to valid predictions. These conditions have the following intuitive interpretations: (i) S_t^y should not depend on the order of data points to preserve the exchangeability of the non-conformity vectors, and (ii) fraction $1-\epsilon$ of the non-conformity vectors should fall into the acceptance set, i.e $|S_t^y \cap \Delta_t^y| \simeq (1-\epsilon)\, t$, to make sure the probability of error is kept at ϵ. These requirements provide a guideline to design acceptance sets. In the next section, we will see that each conformal predictor can be represented in terms of acceptance sets satisfying this conditions. Also we will see an example of acceptance sets tailored for the binary classification problem.

Theorem 3.1 *For a given sequence of* d *dimensional conformity scores* A_t*, acceptance sets* S_t^y*, and smoothing parameters* τ_t*; if for all* $t = 1, 2, \ldots$ *and* $y \in \mathcal{Y}$*:*

i. S_t^y *is measurable conditioned on* $\sigma_{t/y}$,

ii. $|int\,(S_t^y) \cap \Delta_t^y| \leq (1-\epsilon)\,t \leq |S_t^y \cap \Delta_t^y|$,

then the multidimensional conformal predictor associated with these as described in Algorithm 2 *is exactly valid.*

The proof is based on the fact that any smoothed conformal predictor is exactly valid (Appendix of (Shafer & Vovk, 2008)). We simply construct a classical non-conformity score based on a given multidimensional predictor that satisfies the conditions of the theorem, and show that the associated predictors generate exactly the same prediction sets.

Proof: First, consider the acceptance set for label y and time t, S_t^y, and assume it satisfies both of the conditions given in the theorem statement. Then, define the non-conformity score

$$B_t\left(\sigma_{t/y}^{(i)}, z_i\right) = \begin{cases} 2 & \text{if } \mathbf{v}_i^y \notin S_t^y \\ 1 & \text{if } \mathbf{v}_i^y \in \overline{int}\,(S_t^y) \\ i/\,(t+1) & \text{if } \mathbf{v}_i^y \in int\,(S_t^y)\,. \end{cases}$$

Next, we consider three exclusive and exhaustive scenarios to demonstrate the equivalence of the conformal predictor associated with B_t and the multidimensional one. Assuming, $z_t = (x_t, y)$, we calculate the p_t^y values for the conformal predictor associated with B_t in each scenario:

- If $\mathbf{v}_t^y \in int\,(S_t^y)$, then y is included in the prediction set for the multidimensional predictor. Also note that the first inequality of the second condition of the theorem implies:

  $$p_t^y = \frac{\tau_t + t - |int\,(S_t^y) \cap \Delta_t^y|}{t} \geq \epsilon.$$

  Thus y is included for both of the predictors.

- If $\mathbf{v}_t^y \notin S_t^y$, the multidimensional predictor will reject y at time t, and also if we calculate the confidence value of y for the conformal predictor, by the second half of condition ii:

  $$p_t^y = \frac{\tau_t\,(t - |S_t^y \cap \Delta_t^y|)}{t} < \frac{t - |S_t^y \cap \Delta_t^y|}{t} \leq \epsilon.$$

- Lastly, if $\mathbf{v}_t^y \in boun\,(S_t^y)$, the corresponding confidence value becomes:

  $$p_t^y = \frac{t - |S_t^y \cap \Delta_t^y| + \tau_t|\overline{int}\,(S_t^y) \cap \Delta_t^y|}{t},$$

  and this value is greater or equal to ϵ if and only if the second condition on the Line 10 of the Algorithm 2 holds.

Since both predictors behave exactly the same for all of these scenarios, we can declare they are equivalent and since the conformal predictor is valid, the multidimensional one also has to be valid. ■

In addition, we can omit the first half of the second assumed condition, i.e. $|int\,(S_t^y) \cap \Delta_t^y| \leq (1-\epsilon)\,t$, at the cost of achieving conservative validity instead of exact validity.

This follows from the fact that the inequality $|int\,(S_t^y) \cap \Delta_t^y| \leq (1-\epsilon)\,t$ is used only in the first scenario of the proof. In that scenario, the multi-dimensional predictor does not cause an error since the label y is included in the predicted set. However, the conformal predictor may cause an error if the inequality is violated. Thus the multidimensional predictor preserves its (conservative) validity. This argument is summarized in Corollary 3.1.1.

Corollary 3.1.1 *The multidimensional conformal predictor described in Algorithm 2 is valid, if* S_t^y *is* σ_t-*measurable and* $|S_t^y \cap \Delta_t^y| \geq (1-\epsilon)\,t$.

This section has extended the conformal prediction framework to multiple dimensional non-conformity scores and has provided some sufficient conditions to achieve the validity guarantees. However, we haven't touched the issue of "How one should choose acceptance sets to obtain efficient predictions?". The answer to this question depends on the choice of the non-conformity scores which will entail a specific design of acceptance sets. In the next section, we will present a simple choice of acceptance sets for 2-dimensional non-conformity vectors in the binary classification setup that achieves more efficient predictions than the traditional one dimensional conformal predictors under some stability assumptions.

4 CONJUGATE PREDICTION FOR BINARY CLASSIFICATION WITH REJECT OPTION

In this section, we focus on a special case of the confidence prediction problem, where the label space consists of only two elements $\mathcal{Y} = \{0, 1\}$. We propose a simple and effective way of choosing acceptance sets for two dimensional conformal predictors based on any given classical conformal predictor.

As mentioned in the introduction, this problem is equivalent to the scenario of *binary classification with reject option* (Denis & Hebiri, 2015), where at each time point t the predictor either makes a point prediction, i.e. 0 or 1, for y_t or refuses (rejects) to make one, i.e. $\Gamma_t^\epsilon = \{0, 1\}$.In this interpretation, validity implies the probability of error for each prediction is equal to or less than ϵ and efficiency implies the reject option is not used frequently. An asymptotic analysis of error and reject options for this scenario when the traditional conformal prediction is employed can be found at Chapter 3 of (Vovk et al., 2005).

The intuitive idea behind conformal prediction is that a prediction y should be taken if its non-conformity score (monotonic with the probability of error) takes on small values with respect to the non-conformity scores of the other data points. The proposed scheme, which we call *conjugate prediction*, says to make a prediction y if its non-conformity score takes a smaller value than the non-conformity score of the alternative prediction, namely $1-y$.

This approach in some sense tries to find a compromise between the maximum likelihood and conformal prediction. Specifically, it will usually choose the most conforming label, thus enhancing efficiency, while preserving validity by requiring the acceptance set to be large enough to cover at least $1-\epsilon$ fraction of the data points. In the next section, we define conjugate predictors rigorously and show their efficiency. In Section 4.2. we will present the comparison of conjugate and conformal predictors on standard machine learning data sets.

4.1 Conjugate Conformal Prediction

Formally, a conjugate predictor associated with a given one dimensional non-conformity score A_t is a two-dimensional conformal predictor with the non-conformity vectors

$$\mathbf{v}_i^y = \begin{pmatrix} \alpha_i^y \\ \beta_i^y \end{pmatrix} = \begin{pmatrix} A_t\left(\sigma_{t/y}^{(i)}, (x_i, y_i)\right) \\ A_t\left(\sigma_{t/y}^{(i)}, (x_i, 1-y_i)\right) \end{pmatrix},$$

and the acceptance sets

$$S_t^y = \{(\alpha, \beta) : \alpha < \beta \text{ or } \alpha \leq \sup \mathcal{L}_t^y\},$$

where $\mathcal{L}_t^y = \{\gamma : |\{i : \alpha_i^y \leq \gamma \text{ or } \alpha_i < \beta_i^y\}| \leq (1-\epsilon)\, t\}$ and $\sup \emptyset = -\infty$.

For a more intuitive interpretation of the acceptance sets S_t^y, one can imagine to start with the set of points above the $\alpha = \beta$ line (see Figure 1) and combine it with the region $\alpha \leq \gamma$ where γ starts from $-\infty$ and increase the acceptance set until the total number of points in the set equals $(1-\epsilon)\, t$, i.e. $\alpha \leq \sup \mathcal{L}_t^y$, to make sure it satisfies the conditions given in Corollary 3.1.1.

Similarly, the traditional conformal predictor associated with A_t can also be represented as a two dimensional conformal predictor with the same non-conformity vectors $\mathbf{v}_i^y$ and acceptance sets:

$$\tilde{S}_t^y = \{\alpha : \alpha \leq \sup \mathcal{E}_t^y\}$$

where, $\mathcal{E}_t^y = \{\gamma : |\{i : \alpha_i^y \leq \gamma\}| \leq (1-\epsilon)\, t\}$.

As you can see in Figure 1, the traditional conformal predictor satisfies the same intuition as the conjugate predictor: start from $\gamma = -\infty$ and include points in the acceptance set until the number of non-conformity vectors that satisfy $\alpha \leq \gamma$ inequality become equal to $(1-\epsilon)\, t$. Note that this final γ value becomes equal to $\sup \mathcal{E}_t^y$.

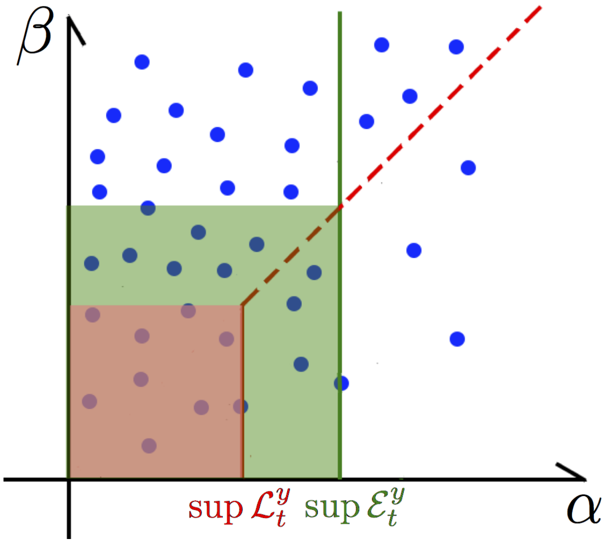

Figure 1: *(A Pictorial View of Conjugate Prediction)* In the figure α-β plane is sketched to illustrate the difference between the conformal and confidence prediction frameworks. For representative purposes, we choose $t = 40$ and $\epsilon = 0.2$. Each non-conformity vector is presented by a blue point assuming $y_t = y$, the acceptance set for conjugate prediction is on the right side of the red line, and the acceptance set for the conformal predictor is on the left of the green line. Assuming the green and red lines are approximately stable (i.e. they do not change based on an individual label value y), the conformal predictor declines to make a prediction if the test point, (x_t, y), falls into either green or red regions, however the conjugate predictor declines only on the red region.

In the following theorem, we argue that a conjugate conformal predictor is more efficient than the conformal predictor associated with the same non-conformity score, if the scoring functions are stable in the following sense. We say a scoring function A_t is stable if it changes little when any single label in the training set flips, i.e. $A_t\left(\sigma_{t/y}^{(i)}, z_i\right) \simeq A_t\left(\sigma_{t/1-y}^{(i)}, z_i\right)$ for any $i < t$. The notion of stability is studied thoroughly in statistical learning theory in the context of necessary conditions for learnability (Shalev-Shwartz, Shamir, Srebro, & Sridharan, 2010) (Bousquet & Elisseeff, 2002). In fact, many of the well-known learning algorithms, as well as the non-conformity scores derived from them, are stable to differing degrees, especially as the number of data points in the training sets increases, i.e. as t increases (Shave-Taylor & Cristianini, 2004), (Bousquet & Elisseeff, 2002), (Shalev-Shwartz & Ben-David, 2002).

In the following theorem, we first present a set of conditions in terms of the auxilary sets $\mathcal{L}_t^y$ and $\mathcal{E}_t^y$ for relative efficiency of the conjugate predictors and then intuitively discuss why the stability of the scoring functions imply these conditions.

Theorem 4.1 *Let scoring functions A_t and tolerance level $\epsilon \in [0,1]$ be given. Also assume the auxilary sets $\mathcal{L}_t^y$ and $\mathcal{E}_t^y$ are defined as:*

$$\begin{aligned} \mathcal{E}_t^y &= \{\gamma : |\{i : \alpha_i^y \leq \gamma\}| \leq (1-\epsilon)\, t\} \textit{ and} \\ \mathcal{L}_t^y &= \{\gamma : |\{i : \alpha_i^y \leq \gamma \textit{ or } \alpha_i^y < \beta_i^y\}| \leq (1-\epsilon)\, t\}. \end{aligned}$$

If $\sup \mathcal{L}_t^y \leq \sup \mathcal{E}_t^{1-y}$ for both $y \in \{0,1\}$, then the conjugate predictor associated with A_t is more efficient than the conformal predictor associated with the same non-conformity score at time t.

Proof: In the proof, we ignore the tie-breaking issues for the sake of brevity, but one can show, with a similar analysis, that the result holds as long as both the conjugate and conformal predictors use the same smoothing value, τ_t.

We start by assuming that the conjugate predictor declines to make a decision at time t, i.e. $|\Gamma_t^\epsilon| = 2$, which implies the non-conformity vector corresponding to the t^{th} data point is included in the acceptance set regardless of the value of y, i.e. $\mathbf{v}_t^y \in S_t^y$ for both $y \in \{0,1\}$.

The definition of the non-conformity vector $\mathbf{v}_t^y = (\alpha_t^y, \beta_t^y)$, implies the equality $\alpha_t^y = \beta_t^{1-y}$ for all $y \in \{0,1\}$. Hence, the conditions for the rejection at time t can be written for any y as:

$$\begin{aligned} \mathbf{v}_t^y &\in \{(\alpha,\beta) : (\alpha < \beta \text{ or } \alpha \leq \sup \mathcal{L}_t^y) \\ &\qquad \text{and } (\alpha < \beta \text{ or } \alpha \leq \sup \mathcal{L}_t^{1-y})\} \\ &\subseteq \{(\alpha,\beta) : \max\{\alpha,\beta\} \leq \max_{y\in\{0,1\}} \sup \mathcal{L}_t^y\}. \end{aligned}$$

To simplify this last statement further, note $\sup \mathcal{L}_t^y \leq \sup \mathcal{E}_t^y$ from the definitions of the auxiliary sets $\mathcal{L}_t^y$ and $\mathcal{E}_t^y$. Combining this inequality with the hypothesis of the theorem, i.e. $\sup \mathcal{L}_t^y \leq \sup \mathcal{E}_t^{1-y}$, we obtain:

$$\max_{y\in\{0,1\}} \sup \mathcal{L}_t^y \leq \min_{y\in\{0,1\}} \sup \mathcal{E}_t^y.$$

Plugging this inequality in the previous statement:

$$\mathbf{v}_t^y \in \{(\alpha,\beta) : \max\{\alpha,\beta\} \leq \min_{y\in\{0,1\}} \sup \mathcal{E}_t^y\},$$

which implies $\mathbf{v}_t^y \in \tilde{S}_t^y$ and $\mathbf{v}_t^y \in \tilde{S}_t^{1-y}$. Therefore, the conformal predictor associated with A_t also declines to make a prediction.

Because the conformal predictor will decline to predict whenever the conjugate predictor will, the conjugate predictor is at least as efficient as the conformal predictor. Further, there are many cases where the conjugate predictor might predict even though the conformal predictor doesn't, for example if the test point fall into the green region in Figure 1. ■

Intuitively, the conditions given in the theorem hold for stable enough non-conformity scores, since stability implies $\sup \mathcal{E}_t^y \simeq \sup \mathcal{E}_t^{1-y}$, and as mentioned before $\sup \mathcal{L}_t^y \leq \sup \mathcal{E}_t^y$ follows directly from the definition of these sets.

The theorem says the conjugate predictor performs at least as efficiently as the original conformal predictor under these stability conditions. The validity of the conjugate predictors follows from Corollary 3.1.1.

In the next subsection, we investigate the relative performance of the conformal and conjugate predictors by comparing the error and rejection rates for different choices of non-conformity scores and datasets. Our main observation is that the conjugate predictors provide two type of gains. First, it reduces the rejection rate due to the extra information provided by the conjugate scores. Second, even if the conjugate score does not provide any extra information (i.e. the β_t^y can be calculated as a function of α_t^y), it reduces the error rate for a given rejection rate, by being more decisive about its choices on easy samples.

4.2 Empirical Results

In this section, we show the results of applying our algorithm and corresponding conformal predictor on some real data-sets from UCI Machine Learning Repository (Lichman, 2013). The details of the used non-conformity scores and the datasets are given in the following two subsections and the numerical results are given at the last subsection.

4.2.1 Experiments

Our experiments use two different non-conformity scores: one based on random forests and the other is based on nearest neighbor classifiers. The reason to choose these two example scores is to illustrate the effect of the conjugate predictor when the conjugate score providing new information about the data (as in the nearest neighbor case), or not (as in the random forests).

1. *Out-of-bag Score in Random Forests:* At each time t, we train a random forest consist of 100 randomized decision trees on $\sigma_{t/y}$. Randomization entails taking a bootstrap of the samples for training and restricting the optimization at the decision nodes to random subsets of the features as described in (Breiman, 2001). We used the Statistics and Machine Learning Toolbox's (MATLAB, 2013) under the default settings, which are the settings suggested by Breiman originally.

   The non-conformity score α_i^y of point z_i is calculated as the fraction of trees (using a training set that doesn't include z_i) that miss-classify the sample x_i, i.e. give the output $1 - y_i$. Note that this choice of non-conformity score implies $\beta_i^y = 1 - \alpha_i^y$, and thus the conjugate score does not provide any new information about the data. Nevertheless, conjugate prediction will still be useful for larger error tolerances.

2. *In-Class Distance in k-Nearest-Neighbor:* As the second scoring function, we built a non-conformity score based on the well-known k nearest neighbor algorithm. For each point z_i, we calculated the closest k data points with the same label y_i from the set $\sigma^{(i)}_{t/y}$ and used the arithmetic average of these k distances as the non-conformity score of α^y_i.

   In the implementation, we tried k values in the range 3 to 10, and we report the results for $k = 5$, which performed the best in all five data sets. Note that while larger values imply better stability, choosing k too large weakens the classifier's predictive power. We used Euclidean distance to measure the closeness of the data points after centering and scaling each feature to unit variance.

   Note that, in contrast to the non-conformity score used with random forests, this non-conformity score provides new information about the data, and as we see in Section 4.2.3, the advantage of using conjugate predictors is greater in this case.

4.2.2 Data Sets

We used the following data sets from UCI ML Repository (Lichman, 2013):

- Breast Cancer Wisconsin (Original) Data Set (Mangasarian & Wolberg, 1990): This data set consists of 699 data points, where each data point is collected from a patient that contains 10 integer valued features of a breast tumor and a binary label for its type (benign/malignant).
- Haberman's Survival Data Set: This data set contains 5 year survival information 306 patients after surgery for breast cancer. The data contain 3 integer features and 1 binary label (survived/died in 5 year.)
- Parkinson's Data Set (Little, McSharry, Hunter, & Ramig, 2008): This dataset contains data about 195 vocal recordings, where each record is represented by a 23 dimensional real vector and the goal is to predict if the subject has Parkinson's disease or not.
- Musk (v1) Data Set: This data set contains 476 data points on 92 types of molecules, each of which is represented by 166 features classifying them as musks and non-musks. The goal is to determine whether a new molecule will be a musk or not.
- Statlog (Australian Credit Approval) Data Set: This data set contains 690 data points on anonymized credit card applications described by 14 features, and classifying them as approved or rejected.

4.2.3 Results

In this part, we tested the above five data sets with both of the described non-conformity scores using error tolerance values of 0.03, 0.10, and 0.18. Because we assume exchangeability, we randomly permute the data before each experiment. Each experiment is repeated 10 times. The means are reported in Table 1,2,3, and 4.

In Tables 1 and 2, we report the cumulative error rate, i.e. fraction of mis-classified samples, for both conjugate and conformal predictors for each score, data, tolerance level combinations. Tables 3 and 4 give the cumulative rejection rates, i.e. the ratio of samples where the classifier declined to predict/classify.

Table 1: *Conjugate Conformal Predictors:* Mean Cumulative Error Rates. KNN means k nearest neighbor, and RF means random forest. B.C. means the breast cancer data set, Surv means survival, and Park. means Parkinson's.

	$\epsilon = 0.03$	$\epsilon = 0.10$	$\epsilon = 0.18$
KNN/B.C.	0.0284	0.0329	0.0335
KNN/Surv.	0.0363	0.1007	0.1699
KNN/Park.	0.0313	0.0836	0.1000
KNN/Musk	0.0336	0.1057	0.1473
KNN/Statlog	0.0358	0.1070	0.1574
RF/B.C.	0.0271	0.0334	0.0334
RF/Surv.	0.0366	0.1000	0.1788
RF/Park.	0.0313	00944.	0.1246
RF/Musk.	0.0372	0.1092	0.1571
RF/Statlog	0.0371	0.1035	0.1380

Table 2: *Classical Conformal Predictors:* Mean Cumulative Error Rates. Labels have the same meaning as in the previous table. Conjugate predictors (previous table) are more accurate or comparable in nearly all cases.

	$\epsilon = 0.03$	$\epsilon = 0.10$	$\epsilon = 0.18$
KNN/B.C.	0.0343	0.1048	0.1827
KNN/Surv.	0.0330	0.0971	0.1683
KNN/Park.	0.0354	0.0979	0.1692
KNN/Musk	0.0368	0.1038	0.1815
KNN/Statlog	0.0371	0.1133	0.1917
RF/B.C.	0.0307	0.1034	0.1892
RF/Surv.	0.0366	0.1000	0.1794
RF/Park.	0.0313	0.0985	0.1697
RF/Musk.	0.0372	0.1092	0.1824
RF/Statlog	0.0371	0.1045	0.1832

Additionally, in Table 5 the cumulative error rates for the native random forest and k nearest neighbor algorithms are presented. For the native implementation, at each time point t, the algorithm is trained on the first $t - 1$ data points and used to predict the t^{th} one. In the for each combination, we report the error rates of the native algorithms over the

Table 3: *Conjugate Conformal Predictors:* Mean Cumulative Rejection Rates. Labels have the same meanings as in the previous tables.

	$\epsilon = 0.03$	$\epsilon = 0.10$	$\epsilon = 0.18$
KNN/B.C.	0.0570	0.0110	0.0098
KNN/Surv.	0.9255	0.7258	0.4507
KNN/Park.	0.4359	0.1410	0.0728
KNN/Musk	0.6149	0.2603	0.0981
KNN/Statlog	0.7581	0.2423	0.0174
RF/B.C.	0.0271	0.0146	0.0143
RF/Surv.	0.7461	0.5020	0.3010
RF/Park.	0.3503	0.1323	0.0805
RF/Musk.	0.4779	0.2088	0.1084
RF/Statlog	0.3964	0.0936	0.0293

Table 4: *Classical Conformal Predictors:* Mean Cumulative Rejection Rates. Labels have the same meanings as in the previous tables. Note that conjugate predictors (previous table) enjoy consistently lower rejection rates for k nearest neighbor algorithm and equivalent rejection rates to the conformal ones upto statistical fluctuations while keeping error rate at a lower level.

	$\epsilon = 0.03$	$\epsilon = 0.10$	$\epsilon = 0.18$
KNN/B.C.	0.6611	0.4246	0.2212
KNN/Surv.	0.9510	0.8013	0.6775
KNN/Park.	0.8195	0.6108	0.4615
KNN/Musk	0.8828	0.6903	0.5221
KNN/Statlog	0.9112	0.6878	0.4467
RF/B.C.	0.0255	0.0102	0.0095
RF/Surv.	0.7461	0.5020	0.3010
RF/Park.	0.3503	0.1297	0.0662
RF/Musk.	0.4779	0.2088	0.0962
RF/Statlog	0.3962	0.0929	0.0148

samples that corresponding conjugate predictors refused to make a prediction or not.

We observe that conjugate predictors always preserve validity (see Table 1), since they reach an error rate equal or less than the target tolerance level (up to statistical fluctuations). However, when the data is relatively easy to classify as in the breast cancer data (see Table 5), conjugate predictors are more decisive while also reducing the error rate by preserving the original validity guarantees.

Furthermore, the decisiveness of conjugate predictors reduces the rejection rates in our simulations (see Table 3 and 4). The gain is more pronounced when the data is relatively less noisy, i.e. easy to classify as in the breast cancer data, and the conjugate score of the base algorithm provides extra information about the data, as when using the k nearest neighbor algorithm.

Table 5: *Baseline:* Depending on the error tolerance ϵ, the conjugate algorithm refuses to predict on certain data points. For each box having format x/y, the table shows the error rate (x) of the underlying algorithm on the refused data points and the error rate (y) on the data points upon which the conjugate algorithm makes prediction. Note that, the error rate is significantly higher on the refused data points whenever the target error level ϵ is low, i.e. refusals are inevitable to preserve the validity.

	$\epsilon = 0.03$	$\epsilon = 0.10$	$\epsilon = 0.18$
KNN/B.C.	0.10/0.03	0.05/0.03	0.00/0.03
KNN/Surv.	0.26/0.54	0.24/0.37	0.23/0.31
KNN/Park.	0.19/0.05	0.21/0.10	0.14/0.11
KNN/Musk	0.23/0.09	0.27/0.14	0.29/0.16
KNN/Statlog	0.16/0.15	0.22/0.14	0.11/0.16
RF/B.C.	0.17/0.03	0.01/0.03	0.00/0.03
RF/Surv.	0.36/0.14	0.41/0.20	0.41/0.26
RF/Park.	0.29/0.05	0.22/0.12	0.06/0.14
RF/Musk.	0.33/0.08	0.37/0.15	0.30/0.19
RF/Statlog	0.26/0.06	0.33/0.12	0.08/0.14

5 DISCUSSION & CONCLUSION

Extending conformal predictors to multiple dimensions is both technically reasonable and practically beneficial. This paper has shown that the extension almost always increases the efficiency and always preserves the validity of machine learning algorithms compared with standard conformal predictors.

Other applications of this extension include scenarios where one may want to combine a set of conformal predictions to make better predictions even when there are breaks in exchangeability. For example, consider the problem of prediction under seasonal changes or other sources of concept drift.

Our conjugate prediction framework is an iterative method for finding hybrid non-conformity scores. As noted in the proof of Theorem 3.1, each multidimensional predictor can be equivalently represented as a conformal predictor. Thus, if one starts with a conformal predictor and can improve upon it by extending it to higher dimensions, as in the case of conjugate predictors, one will obtain a more effective conformal predictor.

The next steps in this work are to demonstrate the benefits of this extension to these other applications, to incorporate the resulting methods into standard machine learning software, and to explore further generalizations of conformal (and multi-dimensional/conjugate conformal) predictors.

References

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2, 499-526.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2), 121-135.

Chaudhuri, K., & Zhang, C. (2013). Improved Algorithms for Confidence-Rated Prediction with Error Guarantees.

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on*, 16(1), 41-46.

Cox, D. R., & Hinkley D.V. (1974). *Theoretical Statistics*. Chapman-Hall, London.

Denis, C., & Hebiri, M. (2015). Confidence Sets for Classification. *Statistical Learning and Data Sciences*, 301-312. Springer International Publishing.

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *The Journal of Machine Learning Research*, 11, 1605-1641.

El-Yaniv, R., & Wiener, Y. (2012). Active learning via perfect selective classification. *The Journal of Machine Learning Research*, 13(1), 255-279.

Herbei, R., & Wegkamp, M. H. (2006). Classification with reject option. *Canadian Journal of Statistics*, 34(4), 709-721.

Kallenberg, O. (2006). *Probabilistic symmetries and invariance principles*. Springer Science & Business Media.

Lei, Jing. (2014). Classification with confidence. *Biometrika*, 101(4), 755-769, doi:10.1093/biomet/asu038.

Lei, J., Robins, J., & Wasserman, L. (2011). Efficient nonparametric conformal prediction regions. *arXiv preprint* arXiv:1111.1418.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 56(4), 1015-1022.

Mangasarian O.L., & Wolberg W.H. (1990). Cancer diagnosis via linear programming. *SIAM News*, 23(5), 1-18, September 1990.

MATLAB and Statistics and Machine Learning Toolbox Release 2013b, The MathWorks, Inc., Natick, Massachusetts, United States.

Meade, N., & Islam, T. (1995). Prediction intervals for growth curve forecasts. *Journal of Forecasting*, 14(5), 413-430.

Ryabko, B., & Ryabko, D. (2013). A confidence-set approach to signal denoising. *Statistical Methodology*, 15, 115-120.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *The Journal of Machine Learning Research*, 9, 371-421.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge University Press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11, 2635-2670.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Vovk, V., Fedorova, V., Nouretdinov, I., & Gammerman, A. (2014). Criteria of efficiency for conformal prediction. *Technical report, Royal Holloway University of London* (April 2014).

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world.* Springer Science & Business Media.