

Aggregating predictions on multiple non-disclosed datasets using conformal prediction

Ola Spjuth^a, Lars Carlsson^b, Niharika Gauraha^a

^a*Uppsala University*

^b*AstraZeneca*

Abstract

The flexible framework for machine learning algorithms called conformal prediction, provides region predictions with guaranteed confidence under mild conditions. Transductive version of conformal predictors have been proven to be valid and more information efficient. In this paper, we propose to aggregate conformal predictions from multiple sources, where transductive conformal predictors are applied on the multiple data sources and their individual predictions are aggregated to form a single prediction on a new example. We illustrate the method using simulated and real data sets, and we show that the proposed method produces much more efficient predictions than individual analysis.

Keywords:

Conformal Prediction, TCP, ACP

1. Introduction

In the biopharmaceutical sciences, it is not unusual for an experiment to be replicated by different manufacturing groups. However, the pooling (or sharing) of experimental data across various manufacturing groups are not encouraged. Also data security is one of the main concerns that has given rise to DataSHIELD approaches (secure analyses that do not require sharing of data). In this article we propose to combine results across experiments without sharing the data, by aggregating conformal predictions computed at individual source level. In particular, we propose to combine conformal p-values from multiple data sources using weighted aggregation or fishers method.

The organization of the paper is as follows. In section 2, we introduce the background concepts and notation, used throughout the paper. In Section 3, we will introduce the concept of aggregating conformal predictions from multiple sources. In Section 4, we discuss the statistical properties of aggregated conformal predictions from multiple sources. In Section 5, we perform some numerical analysis on simulated and real datasets. Finally, in Section 6, the summary of the paper is provided. We have also included an appendix that reviews the most relevant aspects about TCP, ICP, CCP and ACP.

2. Notations and Background

2.1. Set Partition

A partition of a set $S = \{1, \dots, n\}$ is a family of subsets $S_1, \dots, S_k \subset S$, satisfying the following.

1. $S_i \cap S_j = \emptyset$, if $i \neq j$
2. $S = \bigcup_{i=1}^k S_i$
3. $S_i \neq \emptyset$, for $i = 1, \dots, k$

2.2. Transductive Conformity Prediction (TCP)

The object space $\mathcal{X} \in \mathbb{R}^p$, where p is the number of features, and label space $\mathcal{Y} \in (0, 1)$, and example space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ are fixed throughout the article.

The typical classification problem is, given a training dataset $Z = \{z_1, \dots, z_n\}$ – where n is the number of observations in the training set, and each observation $z_i = (x_i, y_i)$ are labeled observations – we want to predict the label of a new observation x whose label is unknown.

First, we define a transductive non-conformity measure and transductive conformity score in the following.

Definition 1 (Transductive nonconformity measure). *A transductive non-conformity measure is a measurable function $\mathcal{A} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ such that $\mathcal{A}(Z_1, Z_2)$ does not depend on the ordering of observations in the set Z_1 .*

Definition 2 (Transductive nonconformity score). *The transductive nonconformity score measures the lack of conformity of the “test observation” Z_2 to the “training set” Z_1 .*

Definition 3 (Transductive Conformity Prediction (TCP)). *Given a training dataset Z and a new observation x , the transductive conformal predictor (TCP), corresponding to a nonconformity measure \mathcal{A} , checks each of a set of hypothesis (for all possible labels) for the new observation x , and assigns it a p -value at a significance level $\epsilon \in (0, 1)$.*

Algorithm 1: TCP

Input: (training dataset: Z , test data: x , label set: \mathcal{Y} , a nonconformity measure: \mathcal{A})

Output: p-values

for each $y \in \mathcal{Y}$ **do**

$z_{n+1} = (x, y)$;

$Z^* = (Z, z_{n+1})$;

 Compute the transductive nonconformity scores:

$\alpha_i = \mathcal{A}(Z^*, z_i)$ for each $z_i \in Z^*$;

 Compute p-value: $p(y) = \frac{|\{i \in \{1, \dots, n+1\} \mid \alpha_i \geq \alpha_{n+1}\}|}{n+1}$;

end

p-values = $\{p(y) \mid y \in \mathcal{Y}\}$;

return p-values;

3. Multi source aggregated TCP Algorithm

Let us consider a binary classification problem, and suppose we have a training dataset Z and external test data set X , or we randomly partition the given dataset into training (80%) and external test set (20%). The algorithm for aggregated TCP from multiple sources is as follows (see Figure 1).

1. The training data set is randomly split into K parts (disjointly) with varying sizes. For example, Let $Z = \{z_1, \dots, z_n\}$ be the data set, then we divide the dataset into S_1, \dots, S_K such that $Z = \bigcup_{i=1}^K S_i$, $k_i = |S_i|$ and $n = k_1 + \dots + k_K$.
2. We compute p-values using (X, S_i) for each S_i , say p_i for each class, then we finally aggregate the k_i p-values (weighted average).
3. We repeat the step 1 and step 2 with different values of K and k_i 's say for q times.
4. Then we analyze the q results obtained (this part is not clear yet).

Algorithm 2: Multi source aggregated TCP

Input: (training dataset: Z , test dataset: X , label set: Y , a nonconformity measure: \mathcal{A})

Output: Aggregated p-values

Initialization;

Unequal size partition: Partition training data randomly and unequally into K parts, S_1, \dots, S_K ;

Steps;

for each $S_i, i \in \{1, \dots, K\}$ **do**

for each observation $x_j \in X$ **do**

 Compute p-values by using **TCP** algorithm:

$PValues_i = \mathbf{TCP}(S_i, x_i, Y, \mathcal{A})$;

end

end

Aggregate $PValues_i$ from various sources into a set **p-values**

return p-values

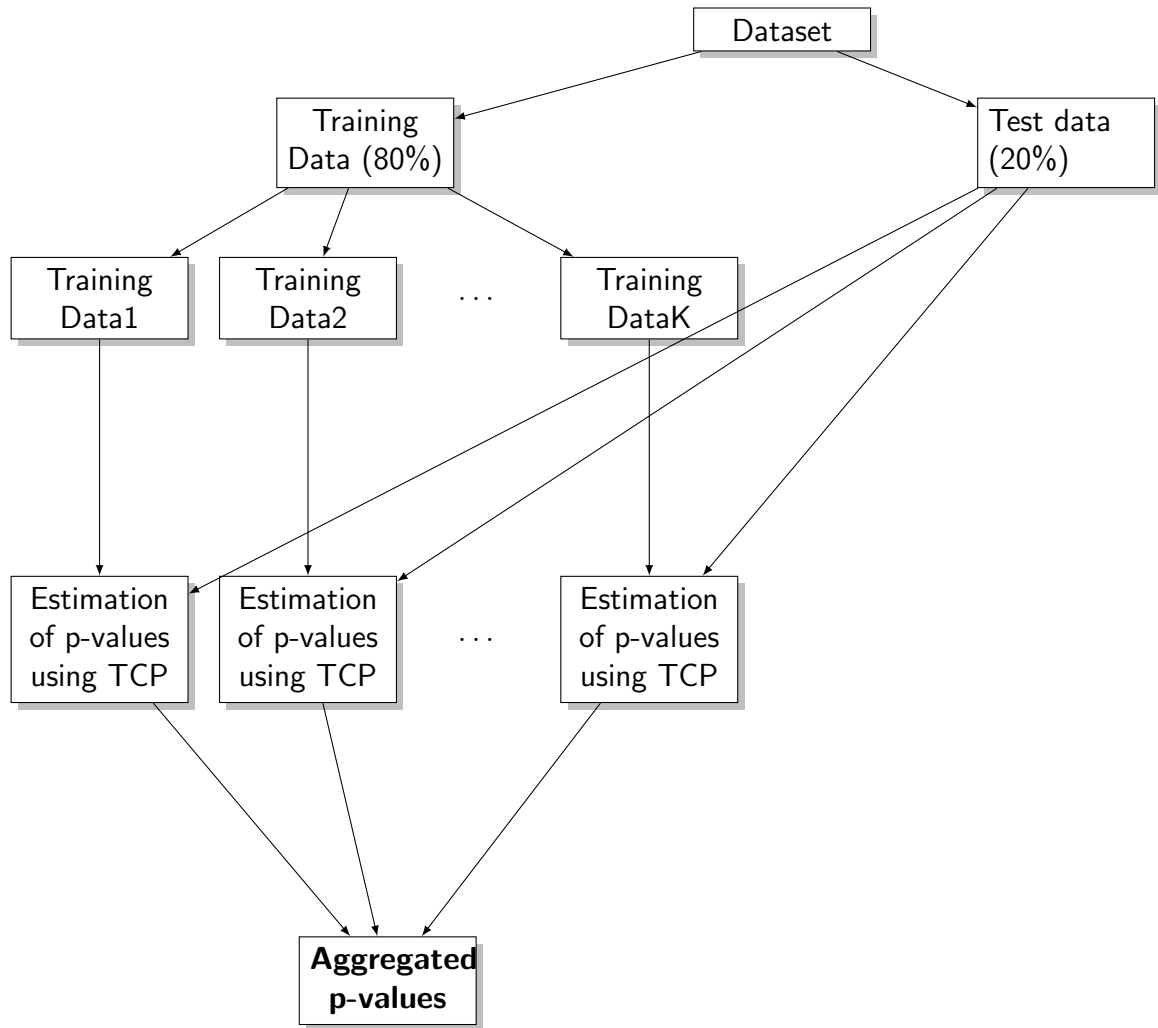


Figure 1: Multi source aggregated TCP Algorithm