

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. In this paper, we extend the basic conformal prediction framework to handle multiple data sources that do not require sharing of data. 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 fisher's method.

Basically, conformal predictors are confidence predictors, that results prediction sets for all confidence levels. Thus, conformal prediction is a framework that complements the predictions of machine learning algorithms with reliable measures of confidence. Transductive version of conformal predictors have been proven to be valid and more information efficient. In this paper, we extend the basic conformal prediction framework to handle multiple data sources and without sharing of data between sources. 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.

The advantages of this approach of combining conformal predictions across multiple sources are two fold. Firstly, it is more a framework than a method, and it extends the existing framework of conformal prediction for multiple data sources, that do not require sharing of data. Secondly, combined analysis produces much more efficient predictions than individual analysis. This innovative framework is flexible in the sense it supports flexible number and sizes of data sources.

At a high level, our algorithm works as follows. Consider a binary classification problem, and suppose we have a training dataset Z and an external test data x . The training data set is randomly and unequally split into K parts. For example, Let $Z = \{z_1, \dots, z_n\}$ be the data set of n observations, then we divide the dataset into S_1, \dots, S_K such that $Z = \bigcup_{i=1}^K S_i$, and $n = k_1 + \dots + k_K$, where $k_i = |S_i|$. We compute p-values under TCP framework using the combined dataset (S_i, x) for each S_i . We have K p-values (for each class), now we aggregate (weighted) the k_i p-values to obtain a final p-value for each class for the new example x . We repeat the process say q times by varying number and sizes of the sources. The final analysis consists of analysing q results which accounts for number as well as size of the data sources.

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.