Review: Combination of Conformal Predictors for Classification

Norio Kosaka

December 18, 2018

1 Paper Profile

- Title: Combination of Conformal Predictors for Classification
- Author: Paolo Toccaceli, Alexander Gammerman
- Organisation: Royal Holloway, University of London, CS Research Centre
- Publish Year: 2017
- URL: http://proceedings.mlr.press/v60/yanovich17a/yanovich17a.pdf

2 Prerequisites

- Conformal Prediction: Intuitively it is the methods that uses past experience to determine precise level of confidence in new predictions, and it employs statistical indicators to evaluate the validity of the prediction Good PPT: http://www.clrc.rhul.ac.uk/copa2017/presentations/CP_Tutorial_2017.pdf

3 Contents in the paper

1. Introduction
2. Results Description
   (a) Manifold Learning Data Model
   (b) Statistics Form
   (c) Main Results
3. Data Model
4. Main Results
5. Proof of Main Theorems
6. Conclusion

4 Abstract

The author proposed some possible approaches to the combination of *Conformal Predictors* in the binary classification case.

- p-value combination techniques
- Calibration of p-values into Bayes factors

And the result shows that the **P-value combined with Fisher's Method** worked fine, when ranking compounds by strength of evidence.

5 Introduction

Conformal Predictors(CP)(Vovk et al., 2005 [6]; Gammerman and Vovk, 2007[3]) was proposed to provide a validity property on the prediction stage.The efficiency of CP hugely relies on underlying *Machine Learning algorithms*. The objective of this paper is to explore ways to improve CP by some forms *ensembling*, which in general refers to the approaches, for example, Random Forests or Bagging. But the author claimed that he has differentiate the proposition from other approaches in a way that it does not explicitly aim at combating overfitting and correlation per se. So, the challenge which he has encountered was to find a method of wide applicability that combines the predictions in a synergistic way.

6 Conformal Predictors

Key points:

- the training set is made up of l independent identically distributed samples
- CP assigns a p-value to a prediction y
- **Non-Conformity Measure(NCM)** is a real-valued function α expressing how odd the sampled example is among the training set.
- Applying one of the following methods to a test set
  - Given a significance level ϵ, a *region predictor* outputs for each test object the set of labels, such that the actual label is not in the set no more than a fraction ϵ of the times.
  - Or, one can pick the largest p-value for a given test object alongside with its credibility and confidence.

Two forms of CP

- Transductive CP: computationally expensive, requires the computation from scratch for each object
- Inductive CP: requires just one training of the underlying, yet also it requires that the training data set which is split into a proper training set and a calibration set.

7 Requirements for CP combination

The research of the problem of combining p-values to obtain a single test for a common hypothesis has a long history, originating to the paper from Fiher's work [2], yet there has not been identified the general enough approach to apply to any kind of real-world problem. Indeed, the combination of p-values from different *CP* on the same test object is a variant of the approaches above and it has aims described below.

- **Preserve Validity**: for the output of the combination method to be a CP, this is a necessary property.
- **Improve Efficiency**: smaller prediction sets must result from a desirable method of combination.

There are two principle in *Mondrian Inductive CP* (http://onlineprediction.net/?n=Main.MondrianConformalPredictor).

1. The p-values from the same CP for the various test objects do not necessarily follow the uniform distribution.
2. The p-values from different CP for the same test object are not independent.

8 Methods from Statistical Hypothesis Testing

In general there are two types of p-value combination methods.

- Order-statistic methods by Davidov(2011)[1]
- Quantile method, here we consider two methods more: Fisher's method and Stouffer's method

8.1 Comparison of Fisher's method(Chi-square method) and Soutffer's method (z-transform method)

"Fisher's method is asymptotically optimal among essentially all methods of combining independent tests" from Littel and Folks(1971) [5]

9 Calibration to Bayes Factors

P-values can be transformed into *Bayes Factors*, which defined as

$$B_\theta(x) = \frac{L_x(\theta)}{\int_\Theta L_x(\theta) dQ(\theta)} \tag{1}$$

where $L_x(\theta)$ is the likelihood of x given θ and $Q(\theta)$ is a prior distribution of θ. So, the interpretation here is that the smaller a Bayes factor is, the less likely it is that the parameter will take value θ having observed data x.

Also a p-value can be transformed into Bayes factor in a way of a *calibrator*, which is that a non-decreasing and continuous function $f : (0, 1) \to (0, +\infty)$ is a calibrator if and only if

$$\int_0^1 \{\frac{1}{f(p)}\} dp \leq 1$$

For instance, a variant of calibrators is given by $f(p) := \frac{p^{1-\alpha}}{\alpha}$ for $\alpha \in (0, 1)$

10 Empirical Results

The authors has tested the three machine learning algorithms combined with NCM using the dataset of the product of a Hihg Throughput Screening assay aimed at identifying chemical compounds that kill cells from a particular tumoral cell line. And the classification into Active vs. Inactive was carried out by applying a threshold on the estimated percentage of cells still alive after exposure to the chemical. And the chosen algorithms are

- Neural network: adapting conforma predictor to the output of the neuron in the output layer.
- SVM: NCM is $-y_i(x_i)$
- Random Forests: NCM chosen for RF was the fraction of trees that classified the test object as having the opposite label as the typothetical one

I would like to refer the readers to the original paper regarding the detailed report of the empirical results.

11 Conclusion

The study demonstrated on a real-world example that, despite their simplicity, these techniques can be of benefit, in particular with the Fisher method exhibiting a synergistic effect on the accuracy of ranking as in the case of the combination of NN and SVM. As for future work, since the tested algorithms only used the bare p-values, we can combine several CP, like *Mixture of Experts models* by (Jacobs et al.,1991 [4])

References

[1] Ori Davidov. "Combining p-values using order-based methods". In: *Computational Statistics & Data Analysis* 55.7 (2011), pp. 2433–2444.

[2] Ronald Aylmer Fisher. "Statistical methods for research workers". In: *Breakthroughs in statistics*. Springer, 1992, pp. 66–70.

[3] Alexander Gammerman and Vladimir Vovk. "Hedging predictions in machine learning". In: *The Computer Journal* 50.2 (2007), pp. 151–163.

[4] Robert A Jacobs et al. "Adaptive mixtures of local experts". In: *Neural computation* 3.1 (1991), pp. 79–87.

[5] Ramon C Littell and J Leroy Folks. "Asymptotic optimality of Fisher's method of combining independent tests". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 802–806.

[6] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Conformal prediction*. Springer, 2005.