Modifications to p-Values of Conformal Predictors

Lars Carlsson[1(✉)], Ernst Ahlberg[1], Henrik Boström[2], Ulf Johansson[3], and Henrik Linusson[3]

[1] Drug Safety and Metabolism, AstraZeneca Innovative Medicines and Early Development, Mölndal, Sweden
{lars.a.carlsson,ernst.ahlberg}@astrazeneca.com
[2] Department of Systems and Computer Sciences, Stockholm University, Stockholm, Sweden
henrik.bostrom@dsv.su.se
[3] School of Business and IT, University of Borås, Borås, Sweden
{ulf.johansson,henrik.linusson}@hb.se

Abstract. The original definition of a p-value in a conformal predictor can sometimes lead to too conservative prediction regions when the number of training or calibration examples is small. The situation can be improved by using a modification to define an approximate p-value. Two modified p-values are presented that converges to the original p-value as the number of training or calibration examples goes to infinity.

Numerical experiments empirically support the use of a p-value we call the interpolated p-value for conformal prediction. The interpolated p-value seems to be producing prediction sets that have an error rate which corresponds well to the prescribed significance level.

1 Introduction

Conformal predictors [6] provide an excellent way of generating hedged predictions. Given a prescribed significance level, ϵ, they make errors when predicting new examples at a rate corresponding to ϵ. The conformal predictor is said to be a valid predictor. This property is attained by predicting sets of possible labels rather than individual labels which is often the case for standard machine-learning methods. However, all conformal predictors are conservatively valid which means that the error rate usually is smaller than the required significance level. Conservative validity leads to less or non-optimal efficiency for a predictor and the ideal situation would be for any conformal predictor to be exactly valid, not only because the relation to efficiency but also because we do not want the predictions to deviate from our expectation with respect to the

This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (IIS11-0053) and the Knowledge Foundation through the project Big Data Analytics by Online Ensemble Learning (20120192).

A. Gammerman et al. (Eds.): SLDS 2015, LNAI 9047, pp. 251–259, 2015.
DOI: 10.1007/978-3-319-17091-6_20

error rate. Reduced efficiency has been observed when using inductive conformal predictors on typical regression datasets from the drug-discovery domain [4]. The observation was that for predictions at low significance levels, prediction ranges sometimes exceeded the observed label range, which could be perceived by users of such predictors as the conformal prediction theory is not very useful. In ongoing work where we use trees for inductive conformal predictions and condition on examples in end leaves of the trees, we observe a similar pattern [5]. In both these examples the problem seems to originate from a limited number of calibration examples leading to an insufficient resolution in p-values.

Given that the exchangeability assumption is fulfilled, the main vehicle for achieving improved efficiency is through improved definitions of the nonconformity scores. This is an area of active research and there are many different ways to construct nonconformity scores through for example using different machine-learning algorithms as part of the definition of nonconformity scores. This is a trial and error process. Another possibility would be to see if there are alternative ways of using the nonconformity scores when computing p-values. We are not aware of any previous approaches and this is why we will attempt to put forward suggestions for alternative definitions of p-values.

The organization of this paper is the following. In the next section we propose another definition of a p-value and also a more generally described p-value. We show that they converge to the p-value suggested in [6] as the number of training examples goes to infinity. Then, in section 3, we empirically study validity and efficiency of all three types of p-values. We conclude the paper in the last section discussing the results.

2 Modifications of the P-value

We are considering a set of examples $\{z_1, \ldots, z_n\}$. A general definition of a smoothed conformal predictor is given in [6], however in this section we will study a non-smoothed p-value or its modifications without loss of generality. The definition of a non-smoothed p-value can be expressed as

$$p = \frac{|\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}|}{n}, \tag{1}$$

and we will refer to this as the *original p-value*. In this case α_n is the nonconformity score of a new example z_n. This example can be predicted using a set predictor

$$\Gamma^{\epsilon}(z_1, \ldots, z_{n-1}) : \{z | p > \epsilon\}, \tag{2}$$

and the nonconformity scores are defined by

$$(\alpha_1, \ldots, \alpha_n) := A(z_1, \ldots, z_n), \tag{3}$$

where A is a nonconformity function producing nonconformity scores for z_i under the exchangeability model. In (1) the nonconformity score α_n, corresponding to the new example z_n, is assumed to belong to the same probability distribution as

the nonconformity scores of the previously observed examples and is included in the distribution when its relative position (based on size) is determined. Since all different α_i that we have observed are random discrete variables, we can define a probability density function

$$f(t) = \sum_{i=1}^{n} \frac{1}{n} \delta(t - \alpha_i), \tag{4}$$

where δ is the Dirac delta function. Furthermore, assume $\alpha_i \in [\alpha_{min}, \alpha_{max}]$ and $\alpha_{min}, \alpha_{max} \in \mathbf{R}$. Then

$$\begin{aligned} F(\alpha_n) := Pr(t \geq \alpha_n) &= \int_{\alpha_n}^{\alpha_{max}} \sum_{i=1}^{n} \frac{1}{n} \delta(t - \alpha_i) dt \\ &= \frac{1}{n} \sum_{i=1}^{n} \big(\theta(\alpha_{max} - \alpha_i) - \theta(\alpha_n - \alpha_i)\big), \end{aligned} \tag{5}$$

where θ is the Heaviside step function. If we assume that all α_i are different and that $\alpha_{min} < \alpha_n < \alpha_{max}$ then (5) becomes

$$F(\alpha_n) = \frac{1}{n} \sum_{i=1}^{n} \big(1 - \theta(\alpha_n - \alpha_i)\big). \tag{6}$$

With $\theta(0) := 0$ we get $F(\alpha_n) \equiv p$ and since all α_i are exchangeable the probability that the set predictor Γ^ϵ will make an error is ϵ as $n \to \infty$.

Let us consider an alternative view where we exclude α_n from the distribution of random discrete variables. After all, given the set of previous observations $\{z_1, \ldots, z_{n-1}\}$ we want to assess how nonconforming z_n is. All examples are still equally probable and we do not change the assumption of exchangeability. The probability density function with z_n excluded is

$$\tilde{f}(t) = \sum_{i=1}^{n-1} \frac{1}{n-1} \delta(t - \alpha_i), \tag{7}$$

corresponding to (5) a *modified p-value* can be defined as

$$\tilde{p} := \tilde{F}(\alpha_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} \big(1 - \theta(\alpha_n - \alpha_i)\big). \tag{8}$$

To see what effect our alternative view has we form the difference between the original and modified p-value

$$\begin{aligned} p - \tilde{p} &= \left(\frac{1}{n} - \frac{1}{n-1}\right) \sum_{i=1}^{n-1} \big(1 - \theta(\alpha_n - \alpha_i)\big) + \frac{1}{n} \\ &= \frac{1 - \tilde{F}(\alpha_n)}{n} \\ &= \frac{1 - \tilde{p}}{n}. \end{aligned} \tag{9}$$

Rearranging this last expression leads to $\tilde{p} = (np - 1)/(n - 1)$ and as $n \to \infty$ the modified p-value converges to the original p-value. Thus, we have shown the following proposition.

Proposition 1. *A conformal predictor based on the p-value in (8) is asymptotically conservatively valid.*

An alternative way of representing the modified p-value is

$$\tilde{p} = \frac{|\{i = 1, \ldots, n-1 : \alpha_i \geq \alpha_n\}|}{n-1}, \tag{10}$$

and smoothing can be applied similarly to how it is applied to the original p-value, since smoothing is a way to ensure that p-values ultimately are distributed uniformly even though there are two or more nonconformity scores that are identical. It is interesting to observe that, given that the nonconformity scores are sorted and renumbered in increasing order, for one particular i then $\alpha_{i-1} < \alpha_n < \alpha_i$ where $i = 2, \ldots, n-1$. For any α_n within these bounds the p-value will be constant. In the following in addition to sorted nonconformity scores, also assume that $0 \leq t \leq 1$ and $t \in \mathbf{R}$. Furthermore, let there exist a monotonically increasing function $g(t)$ with known values $g(t_0 := 0) = \alpha_{min}$, $g(t_{n-1} := 1) = \alpha_{max}$, $g(t_n) = \alpha_n$ and $g(t_i) = \alpha_i$ for $i = 1, \ldots, n-2$. The corresponding probability density function is

$$f_g(t) = \sum_{i=1}^{n-1} g(t)\big(\theta(t - t_{i-1}) - \theta(t - t_i)\big), \tag{11}$$

with a p-value defined as

$$p_g := F_g(t_n) = Pr(t \geq t_n) = \frac{\int_{t_n}^{1} f_g(t)dt}{\int_{0}^{1} f_g(t)dt}. \tag{12}$$

NB we do not have to explicitly compute this expression as long as $\tilde{p} \leq p_g$ and this is obviously true for any α_n and i such that $\alpha_{i-1} < \alpha_n < \alpha_i$ where $i = 1, \ldots, n-1$. Hence, we only need to estimate t_n to get p_g. To conclude this section, we have now shown the following proposition.

Proposition 2. *A conformal predictor based on the p-value in (12) is asymptotically conservatively valid.*

3 Empirical Results of Modified P-values

In this section we will empirically study the validity and efficiency of conformal predictors based on the three different p-values described in the previous section. We are particularly interested in the case of small calibration sets. But first we

will define a p-value corresponding to (12). Given that the nonconformity scores of the already learnt examples are sorted in increasing order with respect to their index, that is $\alpha_i < \alpha_{i+1}$, find a $k = 0, \ldots, n-2$ such that $\alpha_k < \alpha_n < \alpha_{k+1}$. Then an alternative expression to (12) based on determining t_n by linear interpolation using $g(t_k)$, $g(t_{k+1})$ and $g(t_n)$ is,

$$p_g = \frac{|\{i = 1, \ldots, n-1 : \alpha_i \geq \alpha_n\}| - 1}{n-1} + \frac{1 - \frac{\alpha_n - \alpha_k}{\alpha_{k+1} - \alpha_k}}{n-1}. \tag{13}$$

The second term comes from the linear interpolation and we remark that other approximations can be used to more accurately describe $g(t_n)$ as long as they are monotonically increasing. Even for this p-value, smoothing can be applied in exactly the same way as it is applied for the original p-value. We will in the remainder of this paper refer to the p-value in (1) as the *original p-value*, to (10) as the *modified p-value* and finally to (13) as the *interpolated p-value*.

For the numerical experiments, we have used both binary classification responses and regression responses. We used the random forest implementation of the scikit-learn package [1] with default parameters and always 100 trees. The nonconformity scores for classification was defined as

$$\alpha = 1 - \hat{P}(y \mid h(x)), \tag{14}$$

and for regression it was

$$\alpha = \frac{|h(x) - y|}{\sigma(x)}, \tag{15}$$

where $h(x)$ is a prediction of an object x and y the object's label. $\hat{P}$ is what scikit-learn denotes probability. In the case of regression, we trained a second model on log-residual errors and the corresponding prediction of an object is $\sigma(x)$. The prediction sets for classification were calculated using (2) and in the regression case the following expression was used to compute the prediction range for a given significance level ϵ

$$\hat{Y}^{\epsilon} = h(x) \pm \alpha_{\epsilon}\sigma(x), \tag{16}$$

and the midpoint of the range, $y = h(x)$. The datasets used in the empirical study are listed in Table 1. All datasets were taken from UCI [3] except `anacalt`, which was taken from KEEL [2]. The response variables of the regression datasets were linearly scaled to have values in the range $[0, 1]$ and the objects were not scaled. The procedure we followed to generate the empirical results is outlined in Algorithm 1.

Table 1. Datasets

(a) Classification

Dataset	Features
balance-scale	4
diabetes	8
mushroom	121
spambase	57
tic-tac-toe	27

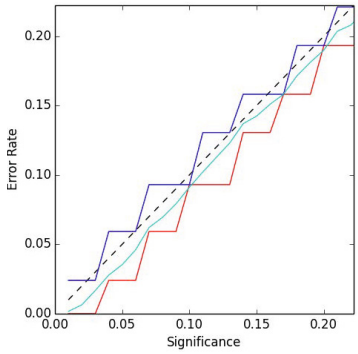
(b) Regression

Dataset	Features
abalone	8
anacalt	7
boston	13
comp	12
stock	9

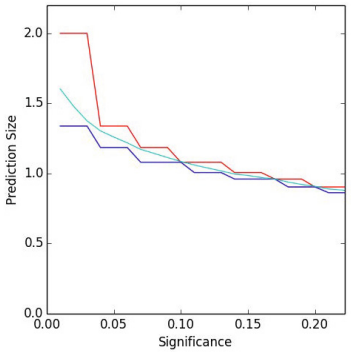
```
for iter ∈ {1, 20} do
    for dataset ∈ datasets do
        for calibrationSize ∈ {29, 49, 99} do
            trainX, testY = dataset.drawrandom(200)
            testX, testY = dataset.drawrandom(100)
            calX, calY = dataset.drawrandom(calibrationSize)
            model = train(trainX, trainY)
            nonconformityScores = calibrate(model, calX, calY)
            predict(testX, model, nonconformityScores, original)
            predict(testX, model, nonconformityScores, modified)
            predict(testX, model, nonconformityScores, interpolated)
        end
    end
end
```

Algorithm 1. This represents how the numerical experiments were conducted. The function `dataset.drawrandom(n)` randomly draws, without replacement, a subset of size n from dataset.

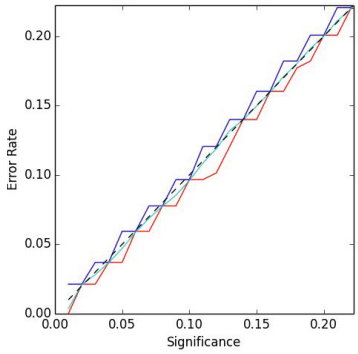
The results for classification and regression were calculated averaging the results for each individual dataset and they are presented in Figures 1 and 2, respectively and are shown for significance levels from 0 to 0.2. The main observation is that for both types of labels, in terms of error rate as well as for efficiency, the interpolated p-value seems to be upper bounded by the original p-value and lower bounded by the modified p-value. Also, the interpolated p-value changes smoothly as opposed to the other two that have a saw-toothed shape. The order of the p-values in terms of increasing efficiency is original, to interpolated and then to modified. In terms of increasing error rate the order is reversed.



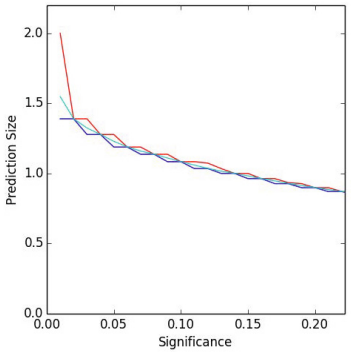
(a) Errors, 29 calib. examples.



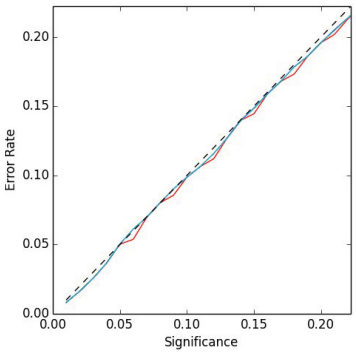
(b) Size, 29 calib. examples.



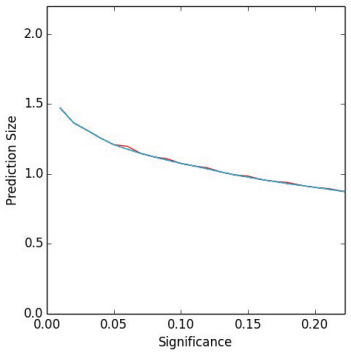
(c) Errors, 49 calib. examples.



(d) Size, 49 calib. examples.

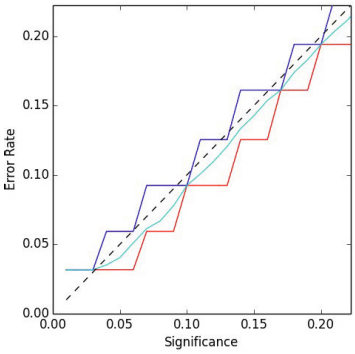


(e) Errors, 99 calib. examples.

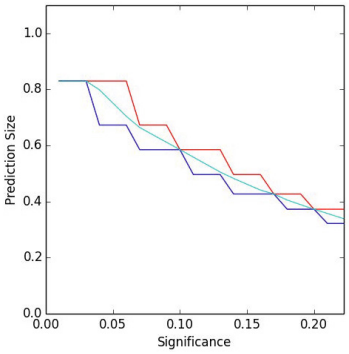


(f) Size, 99 calib. examples.

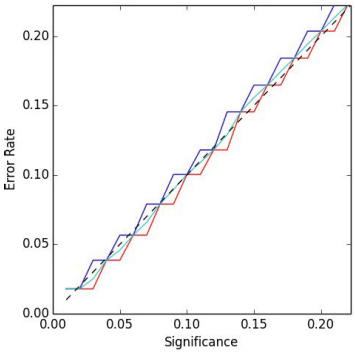
Fig. 1. Classification error rates (left) and prediction sizes (right, average number of classes per prediction) for original p-values (red), modified p-values (blue), interpolated p-values (cyan)



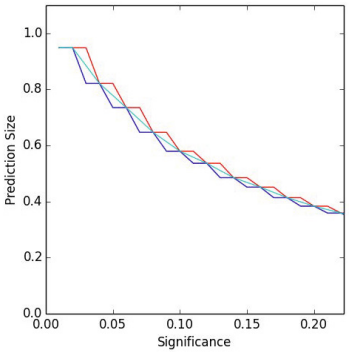
(a) Errors, 29 calib. examples.



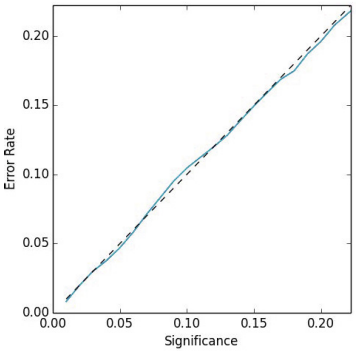
(b) Size, 29 calib. examples.



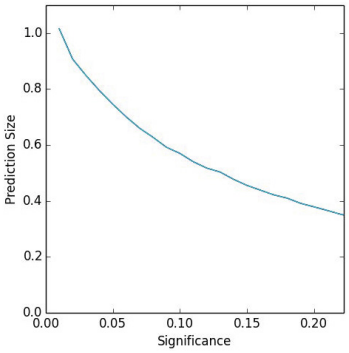
(c) Errors, 49 calib. examples.



(d) Size, 49 calib. examples.



(e) Errors, 99 calib. examples.



(f) Size, 99 calib. examples.

Fig. 2. Regression error rates (left) and prediction sizes (right, average portion of the label space covered per prediction) for original p-values (red), modified p-values (blue), interpolated p-values (cyan)

4 Discussion

We have introduced a new type of p-value which is based on interpolated nonconformity scores. The interpolated p-value is shown to be asymptotically conservatively valid and empirically more close to being exactly valid than the original p-value proposed in [6]. We remark that these differences are only observed when the learnt distribution of calibrations examples are relatively small *i.e.* under 100 examples. For larger samples there does not seem to be a difference and our theoretical analysis show that all p-values covered in this paper converge to the same value for one particular set of examples as the example size goes to infinity. The computational complexity of the interpolated p-value is of the same order as the original p-value. In future work, we will try to compare the original p-value to the interpolated p-value when applied to different datasets and also with different nonconformity functions.

References

1. scikit-learn 0.15 (2014). http://scikitlearn.org/
2. Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing **17**, 255–287 (2010)
3. Bache, K., Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
4. Eklund, M., Norinder, U., Boyer, S., Carlsson, L.: Application of conformal prediction in QSAR. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S. (eds.) AIAI 2012 Workshops. IFIP AICT, vol. 382, pp. 166–175. Springer, Heidelberg (2012)
5. Johansson, U., Ahlberg, E., Boström, H., Carlsson, L., Linusson, H., Sönströd, C.: Handling small calibration sets in mondrian inductive conformal regressors (2014). Submitted to SLDS 2015
6. Vovk, V., Shafer, G., Gammerman, A.: Algorithmic learning in a random world. Springer, New York (2005)