ON THE CALIBRATION OF AGGREGATED CONFORMAL PREDICTORS

COPA 2017 Stockholm

Henrik Linusson[1], Ulf Norinder[2,3], Henrik Boström[3], Ulf Johansson[1,4], Tuve Löfström[1,4]
henrik.linusson@hb.se

June 15, 2017

[1]Department of Information Technology, University of Borås
[2]Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Sweden
[3]Department of Computer and Systems Sciences, Stockholm University
[4]Dept. of Computer Science and Informatics, Jönköping University, Sweden

OVERVIEW

Conformal Prediction

Aggregated Conformal Predictors

Validity of Aggregated Conformal Predictors

CONFORMAL PREDICTION

Conformal classifiers (Vovk et al., 2006) provide us with

Confidence sets	$h(x_i, \epsilon) = \Gamma_i^\epsilon \subseteq Y$
Confidence predictions	$h(x_i) = (\hat{y}_i, \epsilon_y^i)$

Conformal classifiers (Vovk et al., 2006) provide us with

Confidence sets	$h(x_i, \epsilon) = \Gamma_i^\epsilon \subseteq Y$
Confidence predictions	$h(x_i) = (\hat{y}_i, \epsilon_y^i)$

We know the probability of these predictions being correct

$P(y_i \in \Gamma_i^\epsilon) = 1 - \epsilon$
$P(y_i = \hat{y}_i) = 1 - \epsilon_y^i$

- For confidence sets user specifies a (well-calibrated) error rate ϵ
- For confidence predictions the conformal predictor finds the (well-calibrated) error probability of the most likely prediction

Achieved through statistical randomness testing

Divide the training set Z into two disjoint subsets

A proper training set Z_t
A calibration set Z_c where $|Z_c| = q$

Fit a classification model h using Z_t (any learning algorithm will do)

This is the underlying model

Choose an error function $f(z)$, e.g. $f(z_i) = 1 - \hat{P}_h(y_i \mid x_i)$

This is the nonconformity (strangeness) function[1]

Apply $f(Z)$ to $\forall z_i \in Z_c$

Save these calibration scores
We denote these $\alpha_1, ..., \alpha_q$

[1] In reality, the nonconformity function can be any function $f : Z \to \mathbb{R}$, but using (classifier + error function) typically works well for classification problems

Fix a significance level $\epsilon \in (0, 1)$

For each $\tilde{y} \in Y$

Let $\alpha_i^{\tilde{y}} = f[(x_i, \tilde{y})]$

Calculate

$$p_i^{\tilde{y}} = \frac{\left|\left\{z_j \in Z_c : \alpha_j > \alpha_i^{\tilde{y}}\right\}\right|}{q+1} + \theta_i \frac{\left|\left\{z_j \in Z_c : \alpha_j = \alpha_i^{\tilde{y}}\right\}\right| + 1}{q+1}, \theta_i \sim U[0, 1]$$

Prediction region

$$\Gamma_i^{\epsilon} = \left\{\tilde{y} \in Y : p_i^{\tilde{y}} > \epsilon\right\}$$

Transductive conformal predictors are (relatively) slow

Need to retrain underlying model (at least) once for each new test object

Indcutive conformal predictors are (relatively) inefficient

Need to use some of the training data for calibration

Aggregated conformal predictors

- Cross-conformal predictors (Vovk, 2015)
- Bootstrap-conformal predictors (Vovk, 2015)
- Aggregated conformal predictors (Carlsson et al., 2014)

Try to solve both issues:

- All data used for training and calibration
- Model(s) trained only once

AGGREGATED CONFORMAL PREDICTORS

1. Divide training data into k folds $Z_1, \ldots, Z_k$
2. For each fold l
   2.1 Fit a model using $Z_1, \ldots, Z_k \setminus Z_l$
   2.2 Compute (calibration) nonconformity scores for Z_l

Computing a p-value:

$$p^{\tilde{y}}_{n+1} = \frac{\sum_{l=1}^{k} \left[\left| \left\{ z_i \in Z_l : \alpha_{i,l} > \alpha^{\tilde{y}}_{n+1,l} \right\} \right| + \theta_{n+1,l} \left(\left| \left\{ z_i \in Z_l : \alpha_{i,l} = \alpha^{\tilde{y}}_{n+1,l} \right\} \right| \right) \right] + \theta_{n+1}}{n+1}$$

NB

$$p^{\tilde{y}}_{n+1} \approx \bar{p}^{\tilde{y}}_{n+1},$$

where $\bar{p}^{\tilde{y}}_{n+1} = \frac{1}{k} \sum_{l=1}^{k} p^{\tilde{y}}_{n+1,l}$, given that $k \ll n$.

1. From training set Z, draw k samples $Z_1, \ldots, Z_k$, where $\forall Z_i : Z_i \subset Z$
2. For each sample l
   2.1 Fit a model using Z_1
   2.2 Compute (calibration) nonconformity scores for $Z \setminus Z_l$

Computing a p-value:

$$p^{\tilde{y}}_{n+1} = \frac{1}{k} \sum p^{\tilde{y}}_{n+1,l}$$

Aggregated conformal predictors seem great

But... are they valid?

- Vovk (2015) shows that cross-conformal predictors can become invalid under certain circumstances (non-transitive nonconformity functions in leave-one-out conformal predictors)
- Carlsson et al. (2014) show that aggregated conformal predictors requires consistent resampling

Consistent resampling

If $Z^n \subset Z^m \rightarrow \lim_{n,m \rightarrow \mathbf{inf}} f(Z^n) = f(Z^m)$

General validity of aggregated conformal predictors is not clear-cut

Conformal predictors are automatically valid when:

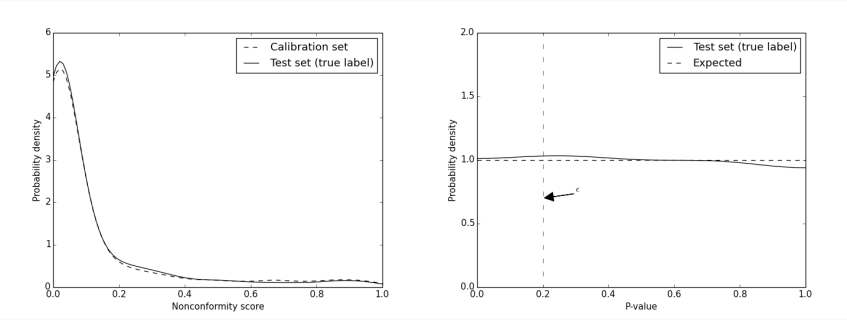
A If a sequence $z_1, \ldots, z_{n+1}$ is exchangeable, then $p_i^{y_i} \sim U[0, 1]$, and

B Criterion **A** is not dependent on the choice of nonconformity function.

Aggregated conformal predictors

- Can fulfill criterion **A** (as shown in several research papers)
- Might not fulfill criterion **B** (?)

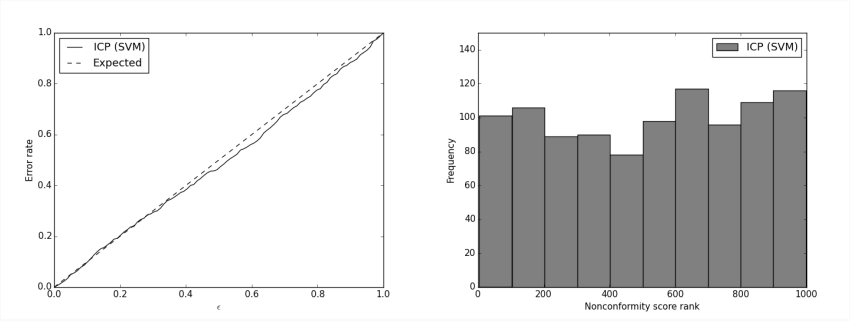
Distribution of nonconformity scores (calibration set and test set, assuming true label) and p-values of test set (assuming true label) for an ICP



(a) Nonconformity distribution

(b) p-value distribution

Calibration plot (error rate per ϵ) and nonconformity-rank distribution of test set (ICP)



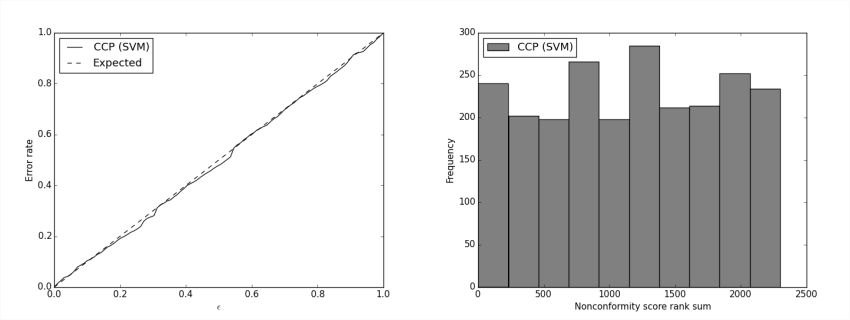
(c) Calibration plot (error rate per ϵ)

(d) Nonconformity rank distribution

Nonconformity rank

A rank r denotes that r − 1 calibration examples had an equal or lower nonconformity score than the test pattern, i.e., it corresponds to the numerator of the p-value equation

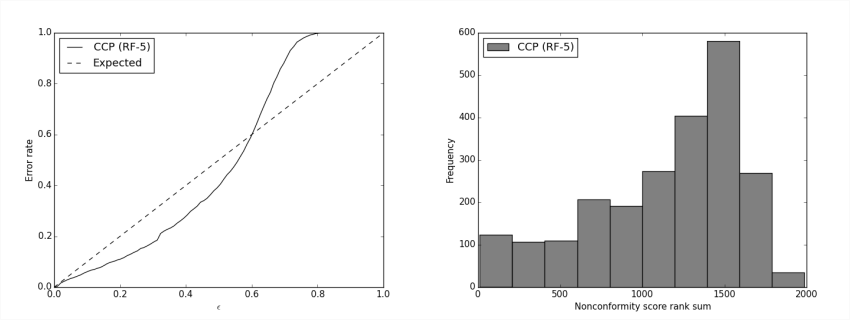
Calibration plot (error rate per ϵ) and nonconformity-rank distribution of test set (CCP using 10 SVM-models)



(e) Calibration plot (error rate per ϵ)

(f) Nonconformity rank distribution

Calibration plot (error rate per ϵ) and nonconformity-rank distribution of test set (CCP using 10 random forest models, each consisting of 5 decision trees)



(g) Calibration plot (error rate per ϵ)

(h) Nonconformity rank distribution

What happened?

VALIDITY OF AGGREGATED CONFORMAL PREDICTORS

In a cross-conformal predictor, we are effectively summing nonconformity ranks (p-value numerators) from several different ICPs

Extreme cases

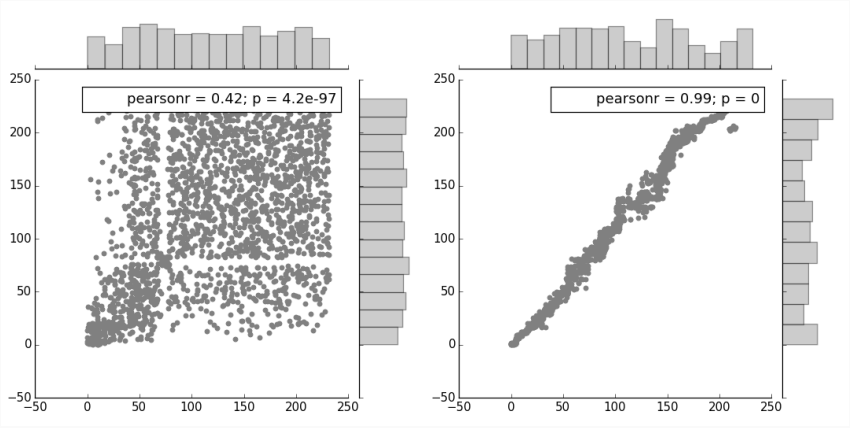
- Ranks from ICP components are identical: no problem (but also no benefit)
- Ranks from ICP components are independent: (?)

Summing (or averaging) multiple independent uniform distributions

Leads to a unimodal Irwin-Hall (or Bates) distribution (uh-oh!)

It seems we have introduced an additional requirement, making validity conditional on the underlying models (or, the nonconformity functions)

Correlation plot of nonconformity ranks for ICP component pairs, (using RF-5 and SVM).
Note: all ICP components produce uniformly distributed ranks separately.



(i) RF using 5 trees → unimodal p-values

(j) SVM → (near-)uniform p-values

So what's the problem?

Random forests with only 5 trees are **unstable** in the sense of Breiman (1996)

- A small change in training data might cause large changes in the resulting classifier
- Hence, each of our 10 random forests (each trained on a separate subset of the avaliable training data) behaves quite differently

Support vector machines are relatively **stable**

- Small changes in training data do not greatly affect the resulting hyperplane
- Hence, each of our 10 SVMs behaves similarly (even though they are trained on separate subsets of the training data)

Unstable classifiers + varied training data $\rightarrow$ loosely correlated nonconformity scores (calibration set and test set) $\rightarrow$ loosely correlated ranks (test set) $\rightarrow$ non-uniform p-values

Approximately valid ACP (informal)

Aggregated conformal predictors are approximately valid when the underlying learning algorithm is stable (produces similar decision boundaries given training data with small variations).

This effectively a restatement of the condition of consistent resampling, but w.r.t to properties of the underlying model.

Approximately valid ACP (informal)

Aggregated conformal predictors are approximately valid when the underlying learning algorithm is stable (produces similar decision boundaries given training data with small variations).

This effectively a restatement of the condition of consistent resampling, but w.r.t to properties of the underlying model.

Approximately valid ACP (formal)

Please read the paper :-)

AVERAGING p-VALUES

Averaging p-values is not a valid approach in general

- We require additional assumptions w.r.t. the dependency of p-values

For interesting (small) values of ϵ, validity is violated in a non-interesting way

- Error of making an error is less than ϵ.

Conservative validity as a result from averaging is problematic primarily w.r.t. efficiency

- Predictions are unecessarily large

Future Work

- Quantify relationship between instability and invalidity
- Quantify impact on efficiency
- Compare to out-of-bag calibration (Boström et al., 2017)
- Consider other methods of combining p-values (briefly evaluated in appendix of presented paper)

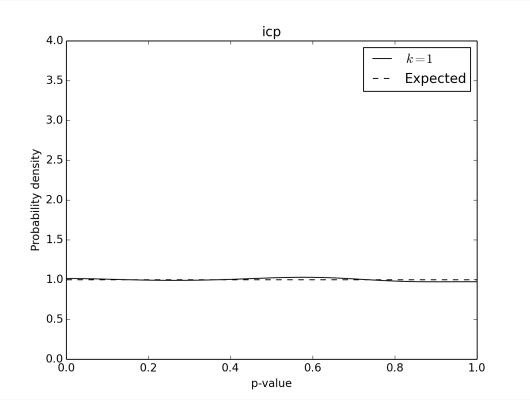
Nonconformist—Conformal Prediction in Python

https://github.com/donlnz/nonconformist

pip install nonconformist

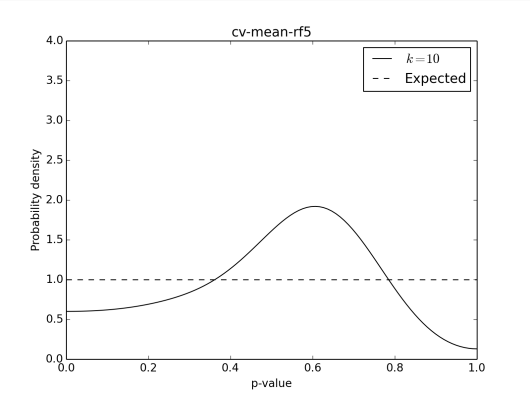
Questions?

Can we do better than averaging?



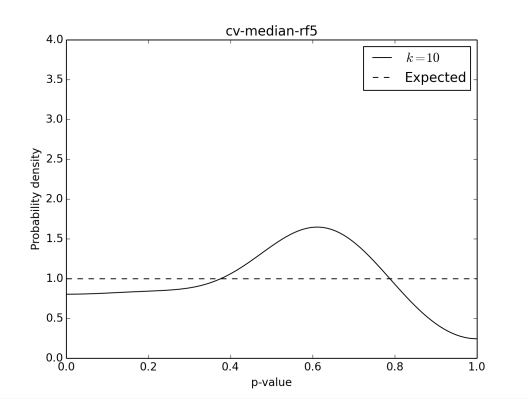
(k) ICP — perfect!

Can we do better than averaging?



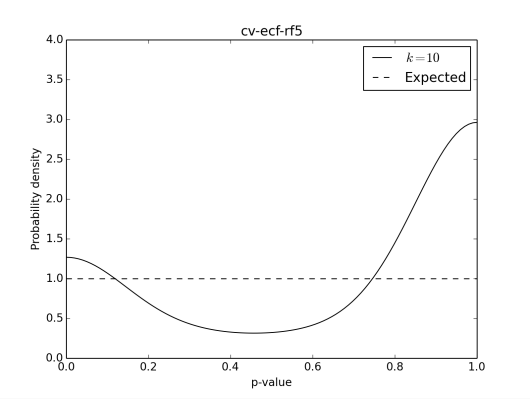
(l) mean(p) — conservative (inefficient) for low ϵ

Can we do better than averaging?



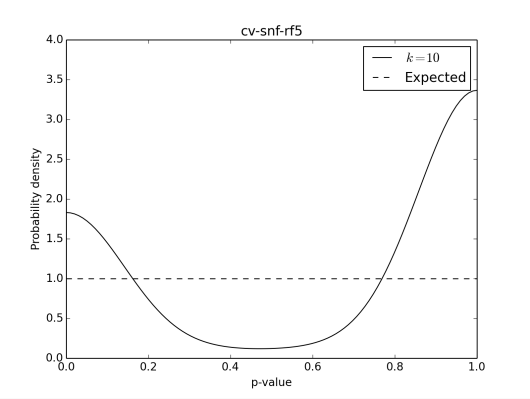
(m) median(p) — empirically conservative (inefficient) for low ϵ, empirically superior to mean(p)

Can we do better than averaging?



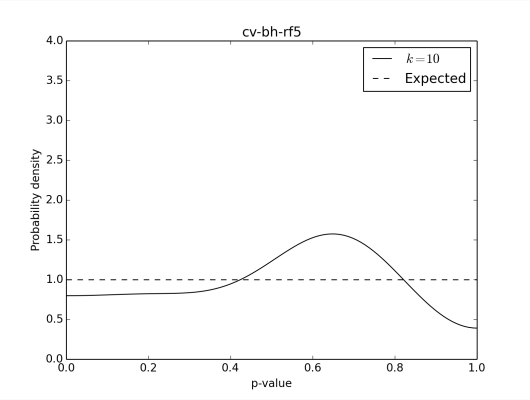
(n) Extended chi-square function, based on Fisher (Balasubramanian et al., 2015) — invalid for low ϵ

Can we do better than averaging?



(o) Simple normal form (Balasubramanian et al., 2015) — invalid for low ϵ

Can we do better than averaging?



(p) Benjamini-Hochberg correction for false discovery rate — empirically conservative (inefficient) for for low ϵ, empirically similar to median(p)

REFERENCES

Balasubramanian, V. N., Chakraborty, S., and Panchanathan, S. (2015). Conformal predictions for information fusion. Annals of Mathematics and Artificial Intelligence, 74(1-2):45–65.

Boström, H., Linusson, H., Löfström, T., and Johansson, U. (2017). Accelerating difficulty estimation for conformal regression forests. Annals of Mathematics and Artificial Intelligence, pages 1–20.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2):123–140.

Carlsson, L., Eklund, M., and Norinder, U. (2014). Aggregated conformal prediction. In IFIP International Conference on Artificial Intelligence Applications and Innovations, pages 231–240. Springer.

Vovk, V. (2015). Cross-conformal predictors. Annals of Mathematics and Artificial Intelligence, 74(1-2):9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2006). Algorithmic learning in a random world. Springer Verlag, DE.