1.5em 0pt

Coverage vs Acceptance-Error Curves for Conformal Classification Models

Evgueni Smirnov

Maastricht University

September 11, 2023

Outline

- Intro to Classification and Conformal Prediction
- Coverage vs Acceptance-Error Graphs for Conformal Predictors
- Coverage vs Acceptance-Error Curves for Conformal Transducers
- Experiments
- Conclusion

Classification Task

Given:

- an object space X,
- a finite set Y of class labels,
- a probability distribution P over $X \times Y$, and
- a training set Tr of M instances $(x_m, y_m) \in X \times Y$ iid drawn from P.

Find:

- a point class estimate $y \in Y$ for test instance $x \in X$ according to P.

Classification Task

In addition,

- test instance x can be supplied by a prediction set $\Gamma(x) \subseteq Y$ that contains possible class labels for $x \in X$ according to P.
- to provide such a set we need a class-label set predictor.

The two most desired properties of a set predictor are:

- **validity:** a class-label set predictor is said to be valid iff the coverage probability that the prediction sets $\Gamma^{\epsilon}(x) \subseteq Y$ do contain the true class labels for test instances x is at least $1 - \epsilon$ for chosen significance level $\epsilon \in (0, 1)$.
- **predictive efficiency:** a class-label set predictor is said to be predictively efficient if the prediction sets $\Gamma^{\epsilon}(x) \subseteq Y$ are non-empty and small.

Conformal Prediction

Given a test instance $x_{M+1} \in X$ and a class label $y \in Y$, the p-value p_y of y for x_{M+1} is computed as follows:

$$p_y = \frac{\#\{(x_m, y_m) \in T | \alpha_m \geq \alpha_{M+1}\}}{M+1} \quad (1)$$

where α_m is a nonconformity score of an instance (x_m, y_m) in $T \cup \{(x_{M+1}, y)\}$ and α_{M+1} is the nonconformity score of (x_{M+1}, y).

Following (Vovk et al., 2016) we consider:

- **conformal predictor** as a set $\Gamma^{\epsilon}(Tr, x_{M+1}) \subseteq Y = \{y \in Y | p_y > \epsilon\}$, where p_y is computed by (1) and ϵ is a significance level in $(0, 1)$, and
- **conformal transducer** as a system $(p_y | y \in Y)$ of p-values over all class labels $y \in Y$.

Criteria for Predictive Efficiency of Conformal Predictors and Transducers (Vovk et al., 2016)

Predictor's Criteria	**Transducer's Criteria**
N	S
M	U
E	F
OM	OU
OE	OF

If Te is a test data set of N instances $(x_n, y_n) \in X \times Y$ iid drawn from P and we fix ϵ, we can estimate **acceptance error rate** AE of conformal predictor Γ^ϵ:

$$AE = \frac{1}{N} \sum_{(x_n, y_n) \in Te} \frac{|\Gamma^\epsilon(x_n) \setminus \{y_n\}|}{|Y| - 1} \tag{2}$$

and **empirical coverage rate** C:

$$C = \frac{1}{N} \sum_{(x_n, y_n) \in Te} \mathbb{1}_{\Gamma^\epsilon(x_n)}(y_n) \tag{3}$$

where $\mathbb{1}$ is the indicator function,

Coverage vs Acceptance-Error Graphs for Conformal Predictors

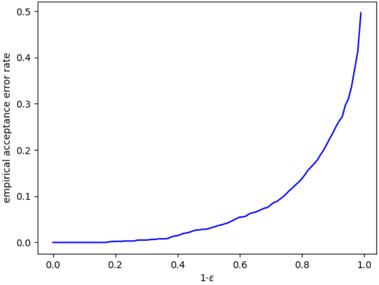


Figure: Empirical acceptance error rate AE in function of $1 - \epsilon$

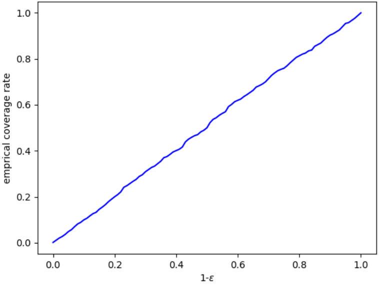
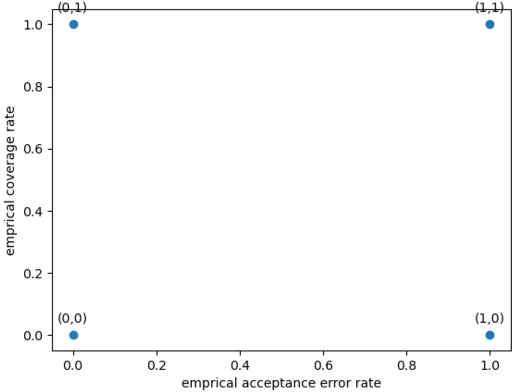


Figure: Empirical coverage rate C in function of $1 - \epsilon$

The Coverage vs Acceptance-Error (CAE) graphs for conformal predictors are two-dimensional graphs in which the empirical coverage rate C is on the Y axis and empirical acceptance error rate AE is on the X axis.



- Conformal predictor Γ_1 dominates conformal predictor Γ_2 iff its empirical coverage rate is greater and its empirical error-acceptance rate is smaller.
- Using predictor dominance we can define the convex hull of the predictor's points in the CAE graph.

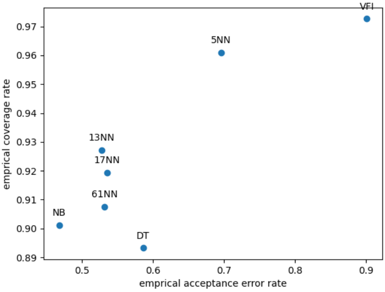


Figure: CAE graph for seven transductive conformal predictors

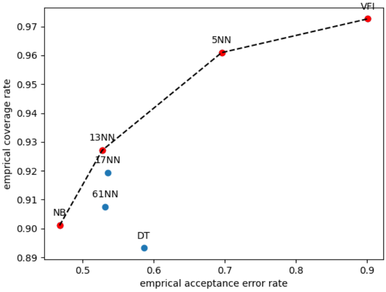
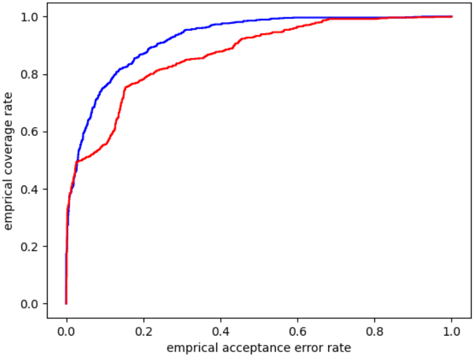


Figure: CAE convex hull for seven transductive conformal predictors

- Assume that we have the system $(p_y | y \in Y)$ for any instance from the test dataset Te.
- We change ϵ from 0 to 1 and plot points (AE, C) of infinitely many predictors Γ^{ϵ} on the CAE graphs.
- These points form Coverage vs Acceptance-Error curves.



Coverage vs Acceptance-Error Curves for Conformal Transducers

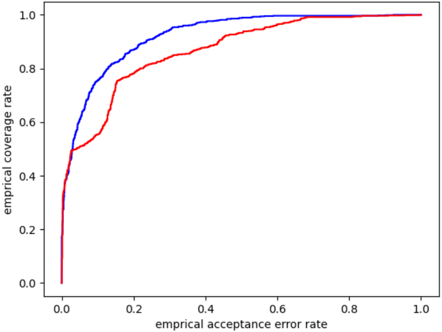
Algorithm 1 Algorithm for Coverage vs Acceptance-Error Curves

Input: List L of triples (n, y, p_y) for each test instance $(x_n, y_n) \in Te$ where p_y is the p-value computed by the conformal class-set predictor for instance x_n and class $y \in Y$.

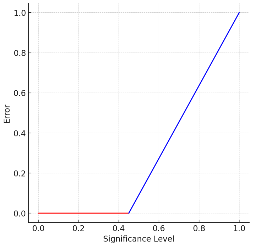
Output: List R of (C, AE) points of the Coverage vs Acceptance Error Curve for list L.

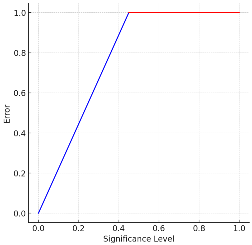
```
Sort the triples (n, y, p_y) in list L in decreasing order of p_y;
R := ∅;
covers := 0;
acceptance_errors := 0;
threshold := +∞;
for next triple (n, y, p_y) ∈ L do
    if p_y ≠ threshold then
        Add tuple (covers/N, acceptance_errors/(N(|Y|−1))) to R;
        threshold := p_y;
    else
        if y = y_n then
            covers := covers + 1;
        else
            acceptance_errors := acceptance_errors + 1;
Output list R.
```

- AUCAEC is the probability that the *p*-value p_{y_n} of randomly chosen true class label y_n of any test instance $(x_n, y_n) \in Te$ is greater than the *p*-value p_y of any other false class label y computed for x_n or any other test instance.
- AUCAEC is label dependent metric. It is independent of significance level and can be used for two-class and multi-class classification tasks.

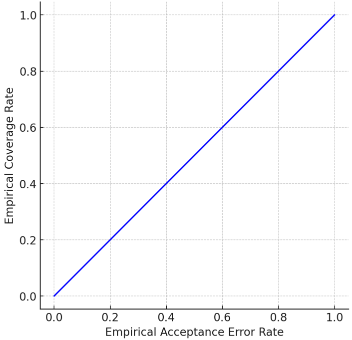


- If AUCAEC $= 1.0$, then p-values p_{y_n} of true class-labels y_n of test instances (x_n, y_n) are greater than the p-values p_y of all the false class-labels y for those instances. Thus, there exist significance levels ϵ for which the empirical coverage rate C is 1.0 and acceptance error rate AE is 0.0; i.e. $\Gamma^{\epsilon}(x_n) = \{y_n\}$ for any test instance (x_n, y_n).
- If AUCAEC $= 0.0$, the p-values p_{y_n} of true class-label y_n of test instances $(x_n, y_n) \in Te$ are smaller than the p-values p_y of all the false class-labels y for those instances. Thus, there exist significance levels ϵ for which the empirical coverage rate C is 0.0 and acceptance error rate AE is 1.0; i.e. $\Gamma^{\epsilon}(x_n) = \{Y \setminus \{y_n\}\}$ for any test instance (x_n, y_n).





- If AUCAEC $= 0.5$ and the CAE curve is close to the diagonal, the probability distribution of the p-values of the true class labels is close to the probability distribution of the p-values of the false class labels.
- Since AUCAEC is in [0,1] from total inefficiency and inaccuracy to total efficiency and accuracy, AUCAEC as a measure for predictive efficiency once the validity has been established.



Experiments

Dataset	AUCAEC	S	U	F	OU	OF
anneal	0.955	0.753	0.123	0.176	0.190	0.248
audiology	0.712	7.365	0.929	6.426	0.937	6.856
autos	0.931	0.971	0.143	0.362	0.227	0.453
balance-scale	0.952	0.651	0.070	0.097	0.081	0.119
breast-w	0.993	0.556	0.004	0.004	0.007	0.007
colic	0.888	0.619	0.060	0.060	0.061	0.115
diabetis	0.804	0.687	0.096	0.096	0.193	0.193
glass	0.952	0.836	0.127	0.246	0.192	0.324
heart-statlog	0.861	0.632	0.070	0.070	0.137	0.137
hepatitis	0.901	0.608	0.055	0.055	0.101	0.101
hypothyroid	0.975	0.580	0.035	0.056	0.055	0.077
ionosphere	0.944	0.562	0.030	0.030	0.058	0.058
iris	0.995	0.525	0.010	0.017	0.012	0.019
lymp	0.891	0.831	0.165	0.211	0.273	0.333
soybean	0.991	0.699	0.024	0.180	0.039	0.195
splice	0.877	0.755	0.119	0.155	0.207	0.246
vehicle	0.923	0.740	0.111	0.151	0.188	0.234
vote	0.979	0.536	0.013	0.013	0.023	0.023
wave	0.897	0.721	0.106	0.114	0.203	0.212
zoo	0.998	0.580	0.015	0.071	0.018	0.074

Experiments: Measures Correlations

Metrics	AUCAEC	S	U	F	OU	OF
AUCAEC	1	0.703	0.767	0.693	0.830	0.706
S	0.703	1	0.983	1	0.935	1
U	0.767	0.983	1	0.979	0.983	0.983
F	0.693	0.999	0.979	1	0.928	0.999
OU	0.830	0.935	0.983	0.928	1	0.936
OF	0.706	1	0.983	0.999	0.936	1

Conclusion

In this paper we have proposed a methodology for visualizing the performances of conformal predictors and transducers.

- **For conformal predictors:** we introduced the coverage vs acceptance-error graphs for visualising the performance of the predictors, their comparison, selection and design on a given significance level ϵ for any k-class classification task for $k \geq 2$.
- **For conformal transducers:** we introduced coverage vs acceptance-error curves. Their area under curve can be viewed as a metric for predictive efficiency if the validity has been established.
- The area under coverage acceptance-curves differs to the existing metrics since it is based on the order of the p-values. It shows the power of the p-values in discriminating class labels.

Questions?

Thank you for your attention!