Coverage vs Acceptance-Error Curves for Conformal Classification Models

Evgueni Smirnov SMIRNOV@MAASTRICHTUNIVERSITY.NL
Maastricht University, Maastricht, The Netherlands

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

In this paper, we introduce coverage vs acceptance-error graphs as a visualization tool for comparing the performance of conformal predictors at a given significance level ϵ for any k-class classification task with $k \geq 2$. We show that by plotting the performance of each predictor for different significance levels in $\epsilon \in [0, 1]$, we receive a coverage vs acceptance-error curve for that predictor. The area under this curve represents the probability that the p-value of randomly chosen true class-label of any test instance is greater than the p-value of any other false class-label for the same or any other test instance. This area can be used as a metric for predictive efficiency of a conformal predictor, when the validity has been established. The new metric is unique in that it is related to the empirical coverage rate, and extensive experiments confirmed its utility and difference from existing predictive efficiency criteria.

Keywords: Conformal Prediction, Predictive Efficiency, Performance Curves and Metrics

1. Introduction

The progress in reliable machine learning in the past two decades has been realized mainly in the context of the conformal prediction framework (Vovk et al., 2005; Shafer and Vovk, 2008; Toccaceli, 2022; Angelopoulos and Bates, 2023). This framework provides a set of techniques for establishing precise level of confidence in new predictions under minimal assumptions on the data. These techniques allow training predictors that output prediction sets with guaranteed coverage; i.e. sets that contain the true value for any new test instance with probability at least $1 - \epsilon$ for a given significance level ϵ. In practice the predictions sets need to be non-empty and small. If this condition holds the conformal predictors are said to be predictively efficient.

One of the interesting problems in conformal prediction is the problem of model selection. So far different metrics for establishing the validity and estimating the predictive efficiency of conformal classifiers have been introduced (Johansson et al., 2013; Vovk et al., 2005, 2016). Techniques for visualizing these metrics have been proposed usually in function of significance level ϵ or of different model parameters. However, there is no metric/visualization that combine simultaneously metrics for the validity and predictive efficiency.

In this paper we introduce a first example of validity vs predictive efficiency graphs for conformal predictors employed in classification tasks. For validity we use the empirical coverage rate while for predictive efficiency we employ the empirical acceptance error rate (defined as the averaged size of the maximal subsets of prediction sets consisting of false class labels). The resulting coverage vs acceptance-error graphs allow visualising the performance of conformal predictors, their comparison, selection and even design on a given significance

level ϵ. When we plot the performance of each predictor for different significance levels $\epsilon \in [0, 1]$, we receive a coverage vs acceptance-error curve for that predictor. Its area under the curve can be viewed as the probability that the p-value of randomly chosen true class-label of any test instance is greater than the p-value of any other false class-label for the same or any other test instance. If the validity has been already established, the area under coverage vs error-acceptance curves can be used as a metric for predictive efficiency.

The rest of the paper is organized as follows. In Section 2 we formalize the classification task. The conformal prediction framework is presented in Section 3. Section 4 provides related work on metrics for predictive efficiency. The coverage vs acceptance-error graphs are introduced in Section 5. In Section 6 we define coverage vs acceptance-error curves. The area under curve is introduced in Section 7. The experiments are provided in Section 8. Section 9 concludes the paper.

2. Classification

Let X be an object space and Y be a finite set of class labels. We assume an unknown probability distribution P defined over $X \times Y$. Training data set Tr is a multi set of M instances $(x_m, y_m) \in X \times Y$ i.i.d. drawn from P. The classification task is to find a point estimate $y \in Y$ of the true class label for a test instance $x \in X$ according to P.

In addition to the point estimate 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 such predictor are validity and predictive efficiency. 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 $\epsilon \in (0, 1)$. 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.

3. Conformal Prediction

Conformal predictor are class-label set predictors that are automatically valid when the data is i.i.d. generated (Vovk et al., 2005; Shafer and Vovk, 2008; Toccaceli, 2022; Angelopoulos and Bates, 2023). They operate as follows. Given a test instance $x_{M+1} \in X$, to decide whether to include a class label $y \in Y$ in prediction set $\Gamma^\epsilon(x_{M+1}) \subseteq Y$, the labeled instance (x_{M+1}, y) is provisionally considered. Then the nonconformity scores α_m of all the instances (x_m, y_m) in $T \cup \{(x_{M+1}, y)\}$ are computed. The p-value p_y of class label y for test instance x_{M+1} is computed as follows:

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

where α_{M+1} is the nonconformity score of (x_{M+1}, y).

Once we have the system of p-values for x_{M+1} computed for all the class labels $y \in Y$ according to (1), we can set the conformal predictor by fixing significance level ϵ. The predictor forms the prediction set $\Gamma^\epsilon(x_{M+1})$ for test instance x_{m+1} from those class labels $y \in Y$ for which $p_y > \epsilon$. In this way we receive validity: the coverage probability that the prediction sets $\Gamma^\epsilon(x_{M+1})$ do include the true class labels is at least $1 - \epsilon$ in a long run.

To compute the prediction sets $\Gamma^{\epsilon}(x_{M+1})$ we need to compute nonconformity scores for the instances in $Tr \cup \{(x_{m+1}, y)\}$ when we assume any class label $y \in Y$ for x_{m+1}. In the transductive conformal predictor a nonconformity score α_m for any instance (x_m, y_m) is determined as a score that indicates how untypical is (x_m, y_m) w.r.t. the instances in data $T \cup \{(x_{M+1}, y)\} \setminus \{(x_m, y_m)\}$. This implies that computing the nonconformity scores is realized by a leave-one-out process that provides superior predictive efficiency for the transductive conformal predictor. However, we note that leave-one-out process is computationally intensive and, thus, the transductive conformal predictor is computationally inefficient.

4. Related Work: p-Value-based Performance Criteria for Predictive Efficiency

We view the proposed area under the coverage vs error-acceptance curve as a metric for predictive efficiency that is independent of significance level. Therefore, we consider here only p-value-based performance criteria for predictive efficiency from (Johansson et al., 2013; Vovk et al., 2005, 2016).

- S-criterion is the average of the sum of p-values p_y for class labels $y \in Y$ over all the test instances.
- U-criterion is the sum of the second largest p-values p_y over all the test instances.
- F-criterion is the average of the sum of p-values p_y for class labels $y \in Y$ minus largest p-values over all the test instances.
- OU-criterion is the sum of the largest p-values p_y for false class labels $y \in Y \setminus \{y_n\}$ over all the test instances (x_n, y_n).
- OF-criterion is the average of the sum of p-values p_y for all the false class labels $y \in Y \setminus \{y_n\}$ over all the test instances (x_n, y_n).

The criteria S, U, and F are label-independent while the criteria OU and OF are label-dependent since the later do employ class-label information from the test instances. Both types of the p-value-based performance criteria indicate predictive efficiency for lower values. However, they may still indicate efficiency for invalid conformal predictors.

5. Coverage vs Acceptance-Error Graphs

Assume that we have a test data set Te defined as a multi set of N instances $(x_n, y_n) \in X \times Y$ i.i.d. drawn from P (just like the training data set Tr). If we fix the significance level ϵ for the conformal predictor, then we can define its empirical coverage rate and acceptance error rate.

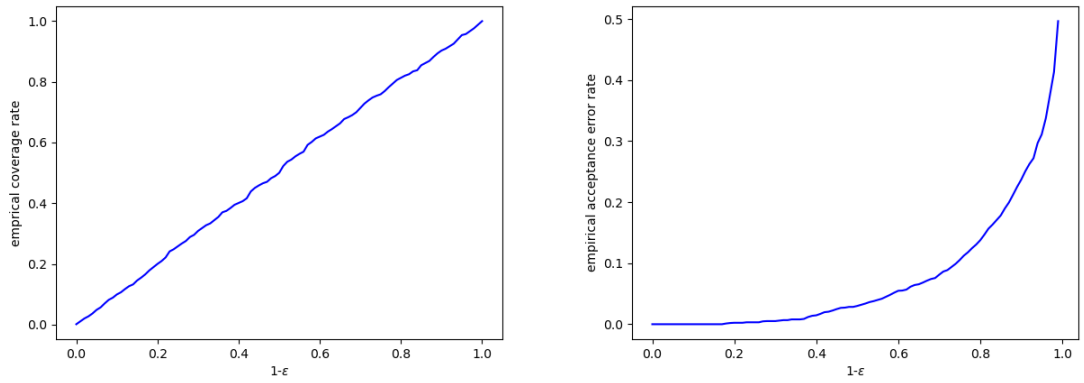
The empirical coverage rate C for a conformal predictor is defined as proportion of the prediction sets $\Gamma^{\epsilon}(x_n)$ that do contain the corresponding true class labels y_n (see Figure 1(a)). More formally,

$$C = \frac{1}{N} \sum_{(x_n, y_n) \in Te} \mathbb{1}_{\Gamma^\epsilon(x_n)}(y_n) \tag{2}$$

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

The empirical acceptance error rate AE is defined as the averaged size of the maximal subsets of prediction sets $\Gamma^\epsilon(x_n)$ that consist of false class labels only (see Figure 1(b)). More formally,

$$AE = \frac{1}{N} \sum_{(x_n, y_n) \in Te} \frac{|\Gamma^\epsilon(x_n) \setminus \{y_n\}|}{|Y| - 1} \tag{3}$$



(a) Empirical coverage rate C in function of confidence $1 - \epsilon$

(b) Empirical acceptance error rate AE in function of confidence $1 - \epsilon$

Figure 1: Emprical rates of transductive conformal predictor based on naive Bayes for the vehicle data (Dua and Graff, 2017). The non-conformity function of this predictor is the general non-conformity function (Vovk et al., 2005).

The Coverage vs Acceptance-Error (CAE) graphs for conformal predictors are two-dimensional graphs in which the empirical coverage rate C is given on the Y axis and empirical acceptance error AE is given on the X axis (see Figure 2). For any significance level ϵ we can visualize the performance of the corresponding conformal predictor by point (AE, C). We note that by construction the CAE graphs can be employed for two-class and multi-class classification tasks.

To employ the CAE graphs we note four special conformal predictors which performance is determined by the following points:

- $(0, 0)$: these conformal predictors output empty prediction sets $\Gamma^\epsilon(x_n)$ for all the test instances $(x_n, y_n) \in Te$. They can be defined for any significance level ϵ that is strictly greater than the maximum of the p-values p_y over all the test instances x_n and class labels $y \in Y$.

- $(1,1)$: these conformal predictors output prediction sets $\Gamma^{\epsilon}(x_n)$ equal to Y for all the test instances $(x_n, y_n) \in Te$. They can be defined for any significance level ϵ that is strictly smaller than the minimum of the p-values p_y over all the test instances x_n and class labels $y \in Y$.

- $(0,1)$: these conformal predictors output prediction set $\Gamma^{\epsilon}(x_n)$ equal to $\{y_n\}$ for any test instance $(x_n, y_n) \in Te$. They can be defined for any significance level ϵ s.t. $(\forall (x_n, y_n) \in Te)((p_{y_n} > \epsilon) \wedge (\forall y \in Y \setminus \{y_n\})(p_y \leq \epsilon))$.

- $(1,0)$: these conformal predictors output prediction sets $\Gamma^{\epsilon}(x_n)$ equal to $Y \setminus \{y_n\}$ for any test instance $(x_n, y_n) \in Te$. They can be defined for any significance level ϵ s.t. $(\forall (x_n, y_n) \in Te)((p_{y_n} \leq \epsilon) \wedge (\forall y \in Y \setminus \{y_n\})(p_y > \epsilon))$.

Special attention deserves conformal predictors which performance is given by points on the diagonal determined by points $(0,0)$ and $(1,1)$. For those predictors the density function of the probability distribution of the p-values of the true class labels is close to the density function of the probability distribution of the p-values of the false class labels.

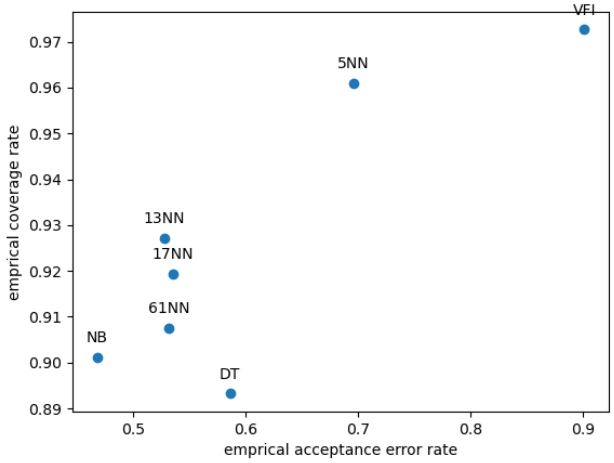


Figure 2: Coverage vs Acceptance Error (CAE) graph for seven transductive conformal predictors based on nearest neighbor (5NN, 13NN, 17NN, 61NN), naive Bayes (NB), decision trees (DT) and voting feature intervals (VFI) tested on the diabetes data (Dua and Graff, 2017) on confidence level of 0.9. The non-conformity function of these predictors is the general non-conformity function (Vovk et al., 2005).

Analyzing the CAE graph in Figure 2 we can define a relationship between conformal predictors (following the approach from the ROC analysis in (Fawcett, 2006)). We can state that conformal classifier h_1 dominates conformal classifier h_2 iff its empirical coverage rate is greater and its empirical acceptance rate is smaller. Given this definition, we can

compute the convex hull of the predictor's points in the CAE graph in Figure 3 [1]. The convex hull has an important property: there is no conformal predictor in the convex hull that dominate any conformal predictor on the convex hull. Thus, any conformal predictor on the convex hull is optimal for certain conditions (expressed in terms of empirical coverage rate and acceptance error rate).

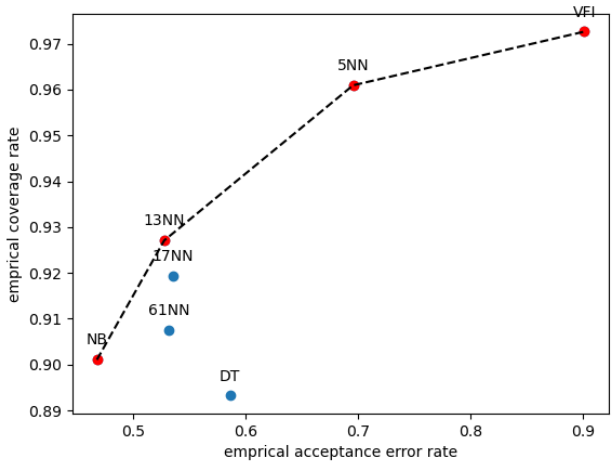


Figure 3: Convex hull in the Coverage vs Acceptance Error (CAE) graph based on seven transductive conformal predictors based on nearest neighbor (5NN, 13NN, 17NN, 61NN), naive Bayes (NB), decision trees (DT) and voting feature intervals (VFI) tested on the diabetes data (Dua and Graff, 2017) on confidence level of 0.9. The non-conformity function of these predictors is the general non-conformity function (Vovk et al., 2005).

In the example in Figure 3 the convex hull is determined by four predictors (plus (0,0) and (1,1) predictors) and we can use any of them. Then the question is whether we can construct any conformal predictor on the convex hull. For that purpose we can easily adapt the procedure proposed by (Fawcett, 2006). Assume that we have two conformal classifiers h_1 and h_2 which points are consecutive on the convex hull. The line segment that connects the points of these predictors has size l and on this segment there is the point of a third conformal predictor h_3 that we wish to achieve. The distance from the point of h_3 to the point of h_1 is l_{13} and the distance from the point of h_3 to the point of h_2 is l_{32}. We receive the performance of conformal predictor h_3 by taking randomly the prediction sets of h_1 for $\frac{l_{32}}{l}100\%$ of test instances and the prediction sets of h_2 for predicting the remaining $\frac{l_{13}}{l}100\%$ of test instances.

1. Computing convex hull can be realized by standard algorithms such as the Jarvis march algorithm (Jarvis, 1973).

6. Coverage vs Acceptance-Error Curves

Assume that we have computed the p-values p_y for all the class labels $y \in Y$ for any instance from the test dataset Te using equation (1). Now if we change the significance level ϵ from 0 to 1, we can in principle receive infinitely many conformal predictors. Assume that we can plot the performance of these predictors using points (AE, C) on the CAE graphs. These points form Coverage vs Acceptance-Error curves (CAE curves for short). They show the tradeoff between the empirical coverage rate C and acceptance error rate AE in function of the significance level ϵ (see Figure 4). We note that by construction the CAE curves can be employed for two-class and multi-class classification tasks.

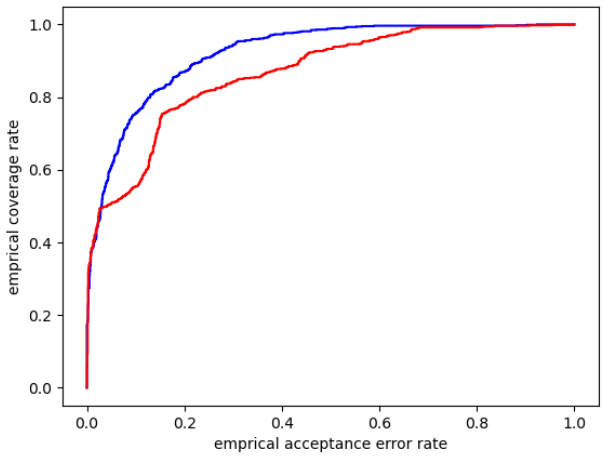


Figure 4: Coverage vs acceptance-error curves for conformal predictor based on naive Bayes (blue) and 5NN (red) for the vehicle data (Dua and Graff, 2017). The non-conformity function of these predictors is the general non-conformity function (Vovk et al., 2005).

The above description provides an intuitive definition of the CAE curves, not an algorithm. We adapt the ROC algorithm of Fawcett (2006) and propose below a computationally efficient algorithm for CAE curves in Algorithm 1. To run the algorithm we first need to form 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 of label $y \in Y$ computed according to formula (1) for x_n. Then we sort the triples (n, y, p_y) in list L in decreasing order of p_y. We set counter *covers* (*acceptance_errors*) initially to 0 since it counts the number of class labels (in)correctly accepted. In addition, we set the *threshold* variable to $+\infty$ since it is used to provide class-label acceptance and initially no class label is supposed to be accepted. Once the counters and *threshold* variable are set, the algorithm visits sequentially the triples (n, y, p_y) in L. If p_y of the current triple (n, y, p_y) differs the value of *threshold*, this is an indication that accepting label y assumes that we need to lower *threshold*. Thus, we preserve the information on what we have accepted so far by adding tuple $(\frac{covers}{N}, \frac{acceptance\_errors}{N(|Y|-1)})$ to the resulting list R of (C, AE)

points (check formulas 2 and 3). If p_y of the current triple (n, y, p_y) equals the value of *threshold*, this is an indication that label y is accepted. If y is the correct label y_n for test instance x_n, then *covers* is incremented. Otherwise, (y is incorrect label for test instance x_n) *acceptance_errors* is incremented. Once all the elements of L have been visited, the algorithm outputs list R of (C, AE) points of the Coverage vs Acceptance Error Curve in the increasing order of AE.

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;
$R := \emptyset$;
$covers := 0$;
$acceptance\_errors := 0$;
$threshold := +\infty$;
for next triple $(n, y, p_y) \in L$ **do**
  if $p_y \neq threshold$ **then**
    Add tuple $(\frac{covers}{N}, \frac{acceptance\_errors}{N(|Y|-1)})$ to R;
    $threshold := p_y$;
  else
    if $y = y_n$ **then**
      $covers := covers + 1$;
    else
      $acceptance\_errors := acceptance\_errors + 1$;
Output list R.

7. Area Under Coverage vs Acceptance-Error Curves

Area Under Coverage vs Acceptance-Error Curves (AUCAEC) is a value in the range of $[0.0, 1.0]$. This is due to the fact that the empirical coverage rate C and the acceptance error rate AE are in the range of $[0.0, 1.0]$. AUCAEC has a straightforward interpretation: it 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. We note that AUCAEC is independent of significance level. Since it employs class-label information, it is label-dependent metric. In addition, AUCAEC can be used for two-class and multi-class classification tasks since it is derived from the CAE curves.

When AUCAEC equals 1.0, the p-values p_{y_n} of true class-label y_n of all the test instances $(x_n, y_n) \in Te$ are greater than the p-values p_y of all the false class-labels y computed for those instances. This implies that 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 $(x_n, y_n) \in Te$. If these significance levels ϵ are greater than 0.0, then the corresponding conformal predictors are conservatively valid.

When AUCAEC equals 0.0, the p-values p_{y_n} of true class-label y_n of all the test instances $(x_n, y_n) \in Te$ are smaller than the p-values p_y of all the remaining class-labels y computed for those instances. This implies that 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 $(x_n, y_n) \in Te$. If these significance levels ϵ are smaller than 1.0, then the corresponding conformal predictors are invalid.

When AUCAEC equals 0.5 and the CAE curve is close to the diagonal $(0, 0) - (1, 1)$, 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. The validity of the conformal predictors can be observed in this case if the distributions are uniform.

Since AUCAEC is in the range of 0 to 1 from total inefficiency and inaccuracy to total efficiency and accuracy, we consider AUCAEC as a measure for predictive efficiency once the validity has been established.

8. Experiments

This section presents our experimental set-up, results, and analysis. We consider twenty data sets provided by the UCI machine learning repository (Dua and Graff, 2017). We experiment with the meta-conformal ensemble proposed in (Smirnov et al., 2006, 2009). It is initialized as follows: the meta predictor that provides conformal prediction is transductive conformal nearest neighbor proposed in (Proedrou et al., 2002). The base classifier is Naive Bayes. The meta-conformal ensembles are tested using a stratified 10-fold cross validation procedure. All of them are found to be valid. Thus, to estimate predictive efficiency we employ the following six metrics: AUCAEC, S Criterion, U Criterion, F Criterion, OU Criterion, and OF Criterion. The results of the experiments are given in Table 1.

In order to analyze the relationship between different predictive-efficiency metrics, we compute the correlation matrix of these metrics based on the data from Table 1. The matrix is given in Table 2.

The matrix in Table 2 provides some evidence that group-wise AUCAEC as a label-dependent metric is more correlated with label-dependent metrics (the OU and OF criteria) than with label-independent metrics (the S, U, and F criteria). The highest correlation of AUCAEC is observed with the U and OU criteria. We note that the U criterion is associated with the second-largest p-values, while the OU criterion is linked to the largest p-values for false class labels. These p-values have a bigger influence on AUCAEC than smaller p-values (since AUCAEC is the probability that the p-value of any true class-label is greater than the p-value of any other false class-label). This explains lower correlation of AUCAEC with the F and OF criteria, respectively.

The correlation matrix in Table 2 shows that all the p-value-based performance criteria (S,U, F, OU, and OF) from Section 4 are highly correlated; i.e. each of them does not bring much extra information than others. In contrast, AUCAEC has the average correlation of 0.73959498 with the p-value-based metrics. This means that AUCAEC is rather different and shows different aspects of the predictive efficiency in relation with the empirical coverage rate.

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

Table 1: Predictive efficiency metric values for 20 UCI data sets for meta transductive conformal ensemble based on meta conformal nearest neighbor and Naive Bayes

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

Table 2: Absolute Pearson correlation coefficient values of different pairs of predictive efficiency metrics based on the data from Table 1

9. Conclusion

In this paper we introduced the coverage vs acceptance-error graphs for visualising the performance of conformal predictors, their comparison, selection and design on a given significance level ϵ for any k-class classification task for $k \geq 2$. When we plotted the performance of these predictors for significance levels $\epsilon \in [0, 1]$, we received coverage vs acceptance-error curves. Their area under curve is viewed as the probability that the p-value of randomly chosen true class-label of any test instance is greater than the p-value of any other false class-label for the same or any other test instance. If the validity has been already established, the area under coverage acceptance-curves can be used as a metric

for predictive efficiency. The distinctive feature of the new metric is that it is related the empirical coverage rate. This confirmed by extensive experiments that showed the utility of the new metric and its difference with the existing efficiency criteria.

We note that the coverage vs acceptance-error curves have some resemblance with the ROC curves for conformal predictors proposed in (Vovk, 2013). However, a key distinction lies in the fact that the ROC space is defined by per-class error rates, limiting its applicability to two-class classification tasks.

Future research will focus on detailed investigation of the properties of the coverage vs acceptance-error graphs and their application for validating conformal predictors in the context of multi-label classification and regression.

Acknowledgments

I would like to express my gratitude to the anonymous reviewers for their valuable feedback and insightful comments on the paper.

References

Anastasios Angelopoulos and Steven Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16:494–591, 2023.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.10.010. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

R. A. Jarvis. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2:18–21, 1973.

Ulf Johansson, Rikard König, Tuve Löfström, and Henrik Boström. Evolved decision trees as conformal predictors. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2013, Cancun, Mexico, June 20-23, 2013*, pages 1794–1801. IEEE, 2013. doi: 10.1109/CEC.2013.6557778. URL https://doi.org/10.1109/CEC.2013.6557778.

Kostas Proedrou, Ilia Nouretdinov, Volodya Vovk, and Alexander Gammerman. Transductive confidence machines for pattern recognition. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002, 13th European Conference on Machine Learning, Helsinki, Finland, August 19-23, 2002, Proceedings*, volume 2430 of *Lecture Notes in Computer Science*, pages 381–390. Springer, 2002. doi: 10.1007/3-540-36755-1_32. URL https://doi.org/10.1007/3-540-36755-1_32.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Evgueni N. Smirnov, Stijn Vanderlooy, and Ida G. Sprinkhuizen-Kuyper. Meta-typicalness approach to reliable classification. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *ECAI 2006, 17th European Conference on Artificial Intelligence, August 29 - September 1, 2006, Riva del Garda, Italy, Including Prestigious Applications of Intelligent Systems (PAIS 2006), Proceedings*, volume 141 of *Frontiers in Artificial Intelligence and Applications*, pages 811–812. IOS Press, 2006.

Evgueni N. Smirnov, Georgi I. Nalbantov, and A. M. Kaptein. Meta-conformity approach to reliable classification. *Intell. Data Anal.*, 13(6):901–915, 2009. doi: 10.3233/IDA-2009-0400. URL https://doi.org/10.3233/IDA-2009-0400.

Paolo Toccaceli. Introduction to conformal predictors. *Pattern Recognit.*, 124:108–507, 2022. doi: 10.1016/j.patcog.2021.108507. URL https://doi.org/10.1016/j.patcog.2021.108507.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92 (2-3):349–376, 2013. doi: 10.1007/s10994-013-5355-6. URL https://doi.org/10.1007/s10994-013-5355-6.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In Alexander Gammerman, Zhiyuan Luo, Jesús Vega, and Vladimir Vovk, editors, *Conformal and Probabilistic Prediction with Applications - 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings*, volume 9653 of *Lecture Notes in Computer Science*, pages 23–39. Springer, 2016. doi: 10.1007/978-3-319-33395-3_2. URL https://doi.org/10.1007/978-3-319-33395-3_2.