Proceedings of Machine Learning Research 204:1–15, 2023 Conformal and Probabilistic Prediction with Applications

A Review of Nonconformity Measures for Conformal Prediction in Regression

Yuko Kato Y.KATO@TUDELFT.NL
David M.J. Tax D.M.J.TAX@TUDELFT.NL
Delft University of Technology, Delft, The Netherlands
Marco Loog MARCO.LOOG@RU.NL
Radboud University, Nijmegen, The Netherlands

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

Abstract

Conformal prediction provides distribution-free uncertainty quantification under minimal assumptions. An important ingredient in conformal prediction is the so-called nonconformity measure, which quantifies how the test sample differs from the rest of the data. In this paper, existing nonconformity measures from the current literature are collected and their underlying ideas are analyzed. Furthermore, the influence of different factors on the performance of conformal prediction are pointed out by focusing on the relation between the influencing factors and the choice of nonconformity measures. Lastly, we provide suggestions for future work with regard to currently existing knowledge gaps and development of new nonconformity measures.

Keywords: Conformal prediction, Uncertainty quantification, Nonconformity measure

1. Introduction

With the increasing popularity of machine learning, concerns regarding its prediction accuracy have risen. Conformal prediction is currently the only method that can provide distribution-free uncertainty quantification (Vovk et al., 2005). Conformal prediction can be used in combination with any machine learning method − including Bayesian methods − and can provide each prediction with a valid confidence measure (Papadopoulos et al., 2011; Fontana et al., 2023). This means that the conformal prediction interval covers the true response with a pre-specified probability. The only required assumption for conformal prediction is that data are exchangeable, unlike other probabilistic machine learning methods that require stronger assumptions on output distributions.

An important aspect of conformal prediction is the so-called nonconformity measure, defining how the test sample differs from the rest of the data (Shafer and Vovk, 2007). In principle, any function can be used as a nonconformity measure. However, it is most likely that the usefulness of the predictive confidence highly depends on the choice of the nonconformity measure (Angelopoulos and Bates, 2021; Papadopoulos, 2008). Therefore, recent work started to focus on the development of new nonconformity measures (Jung et al., 2022). As Papadopoulous and colleagues mentioned, many different nonconformity measures can be constructed for each method and each of these measures potentially defines a different variant of conformal prediction (Papadopoulos et al., 2011). In this paper, we performed a literature review focusing on existing nonconformity measures in the regression setting. These nonconformity measures are assessed and analyzed based on their underlying ideas.

1.1. Outline

This paper is organized as follows; A brief background on conformal prediction is provided in Section 2, where we introduce the two most widely used types of conformal predictors, namely transductive and inductive conformal predictors (Section 2.1). Furthermore, the two main characteristics describing the performance of conformal predictions, validity and efficiency, are described (Section 2.2). In Section 3, we analyze existing nonconformity measures by assessing how they work, motivation to use them and possible limitations. Subsequently, the influences of different factors on the performance of conformal prediction are pointed out in Section 4. In particular, the relation between these influencing factors and the choice of nonconformity measures is highlighted. Furthermore, we discuss considerations regarding a choice of nonconformity measures and suggestions for future work in Section 5. Findings are summarized and the main conclusion is provided in Section 6.

2. Conformal prediction and uncertainty

In order to start the discussion on conformal prediction, let us consider a regression setting using a training dataset of n observations $D = (x_i, y_i)$ with $i \in \{1, \ldots, n\}$. We assume that all the samples $(x_i, y_i)_{i=1}^{n}$ are drawn *i.i.d.* from a joint distribution p_{xy}. Using some underlying method that can make a prediction $\hat{y}$, the aim is to obtain uncertainty information about the unknown value of y_{n+1} at a test point x_{n+1}.

Conformal prediction constructs a marginal distribution-free prediction interval for the test point x_{n+1}, $C(x_{n+1}) \subset R$, based on the training dataset which contains the unknown response y_{n+1}.

$$P\{y_{n+1} \in C(x_{n+1})\} \geq 1 - \epsilon$$

where $1 - \epsilon \in (0, 1)$ is the coverage rate.

Although this holds for any sample size n without any additional assumptions on the underlying data distribution (Tibshirani et al., 2019), it is important to emphasize that this is a marginal coverage. This means that the probability that C satisfies the true test value y_{n+1} is at least $1 - \epsilon$, on average over a random sample from the training and test data.

In order to produce prediction intervals $C(x_{n+1})$, conformal prediction uses a nonconformity measure. This can for instance be a real-valued function that measures the disagreement between the prediction output $\hat{y}_i$, and the actual value y_i, for a given instance x_i (Shafer and Vovk, 2007; Vovk et al., 2005). After computing the nonconformity scores $\alpha_i, (i = 1, \ldots, n)$, which are essentially the output of the nonconformity measure, the nonconformity score at test point x_{n+1}, α_{n+1}, is ranked by computing $\pi(y)$ as follows:

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{\alpha_i \leq \alpha_{n+1}\} = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}\{\alpha_i \leq \alpha_{n+1}\} \tag{1}$$

$\pi(y)$ shows the proportion of points which α_i is smaller than α_{n+1}. Here $\mathbf{1}\{\cdot\}$ is the indicator function. Conformal prediction interval at x_{n+1}, $C(x_{n+1})$, can be obtained using the obtained $\pi(y)$ as follows:

$$C(x_{n+1}) = \{y \in R : (n+1)\pi(y) \leq \lceil (1-\epsilon)(n+1) \rceil\} \tag{2}$$

In this paper, we mainly consider nonconformity measures in a regression setting. In a number of cases, depending on the context of the task, a conformity measure, instead of nonconformity measure can be chosen as well (Fontana et al., 2023). The connection between nonconformity and conformity measures is, however, often fairly simple (e.g. conformity $\equiv 1 -$ nonconformity or conformity $\equiv -$nonconformity). For those interested in such conformity measures instead, see e.g., Vovk (2015); Chernozhukov et al. (2021); Lei et al. (2015).

2.1. Conformal predictors

Conformal prediction methods can be divided into two main categories: transductive conformal predictors (TCP) (Saunders et al., 1999) and inductive conformal predictors (ICP) (Papadopoulos et al., 2002; Papadopoulos, 2008). The main differences between TCP and ICP are their usage of the training data and overall training procedure. In TCP, the whole training dataset is used to train the underlying model. In order to establish the nonconformity baseline, TCP measures the nonconformity of all examples, including $(x_{i+1}, \hat{y}_{n+1})$. This means that the predicted test point $(x_{i+1}, \hat{y}_{n+1})$ is considered when calculating the nonconformity scores for the training set. Therefore, for each new test point, the underlying model needs to be retrained, and the nonconformity measure for the training data has to be recomputed (Linusson et al., 2014). This can be computationally expensive, and this computational burden of TCP can be problematic, especially when the underlying method requires data preprocessing or hyper-parameter selection (Vovk et al., 2022).

In contrast to TCP, ICP first separates the training set into a *proper training set* (to train the underlying algorithm) and a *calibration set* (to obtain the nonconformity scores). Since these two datasets are completely disjoint, the calibration dataset provides unbiased nonconformity measures. This also means that the underlying algorithm has to be trained only once, which makes ICP computationally more efficient.

2.2. Validity and Efficiency

In conformal prediction, two main characteristics (*validity* and *efficiency*) determine the quality of the obtained prediction intervals. A conformal prediction interval is valid when the interval covers the true response y_{n+1} with a predetermined coverage $1 - \epsilon$. In fact, the following proposition holds:

Proposition 1 (Vovk (2012)) *Let Γ be a conformal predictor or an inductive conformal predictor. If random examples $Z_1, ..., Z_i, Z = (X, Y)$ are exchangeable (i.e., their distribution is invariant under their permutations), the probability of error $Y \notin (Z_1, ..., Z_i, X)$ does not exceed ϵ for any ϵ*

Proposition 1 automatically provides a conformal prediction interval of unconditional validity (Vovk, 2012). Furthermore, Vovk mentioned that, in general, a conformal predictor is conservatively valid, meaning that the probabilities for errors are allowed to be even less than ϵ (Vovk et al., 2005).

In a regression setting, another measure indicating the quality of the prediction intervals is the average length of the prediction intervals. This is known as the efficiency and is defined

as follows (Johansson et al., 2013):

$$\frac{1}{n}\sum_{i=1}^{n}|C_i|$$

This is named the *N criterion* by Vovk et al. (2016). Under this criterion, high efficiency is obtained by forming tight prediction intervals. However, this also requires careful consideration when choosing nonconformity measures given that a poor choice of nonconformity measures inherently leads to a decrease in efficiency as explained in Papadopoulos (2008). Other possible efficiency measures are discussed in (Vovk et al., 2016).

In order to get high-quality conformal prediction, we would like to retain validity and maximize efficiency. Much research has focused on making conformal prediction more efficient, see e.g., Linusson (2021); Lei et al. (2013). Although both TCP and ICP are assured to reach the targeted validity, TCP tends to produce prediction intervals with higher validity comparing to ICP (Papadopoulos et al., 2002).

3. Nonconformity measures

Two major categories of nonconformity measures, absolute error-based nonconformity measures and quantile-based nonconformity measures, together with their advantages and disadvantages, are discussed in what follows.

3.1. Absolute error-based nonconformity measures

In a regression setting, the absolute error-based nonconformity measure is the most straightforward nonconformity measure and is defined as:

$$\alpha_i = |y_i - \hat{y}_i| \tag{3}$$

which is the absolute value of the difference between the prediction $\hat{y}_i$ for x_i and y_i (Papadopoulos et al., 2011). We refer to this as the absolute error-based nonconformity measure.

It is important to note that the absolute error-based nonconformity measure, as defined in Equation (3), provides prediction intervals (using Equation (2)) that have the same length for all test examples, and thus potentially affects the efficiency of conformal prediction. Although the fact that intervals have the same width for all test samples pose no difficulties when the research focuses on the validity of conformal prediction (e.g., Vovk (2012)), it can have a drawback since efficiency of conformal prediction can be sacrificed in some cases. This can be for instance the case when data noise is heteroscedastic, where the average length of intervals increase, resulting in a lower efficiency.

To overcome the aforementioned issue, the normalized nonconformity measure (often referred to as locally weighted nonconformity measure) was proposed. This is obtained through an adjustment of the first measure in Equation (3). The normalized nonconformity measure was defined as follows:

$$\alpha_i = \left|\frac{y_i - \hat{y}_i}{\sigma_i}\right| \tag{4}$$

where $\sigma_i = e^{\mu_i}$, μ_i is the prediction of the value $\ln(|y_i - \hat{y}_i|)$. Subsequently, the prediction interval for x_i is obtained as follows:

$$C(x_i) = (\hat{y}_i - \alpha_{s(\epsilon)}\sigma_i,\ \hat{y}_i + \alpha_{s(\epsilon)}\sigma_i) \tag{5}$$

where $s(\epsilon) = \lfloor \epsilon(n+1) \rfloor$. The logarithmic scale is used to ensure that the estimate is always positive. After training the underlying model on training data, we calculate the residuals $|y_i - \hat{y}_i|$ for all training examples $i = 1, ..., n$. Thereafter, the underlying model is retrained on the pairs $(x_i,\ \ln(|y_i - \hat{y}_i|)$. These predicted values are used to obtain the normalized nonconformity measure. It should be noted that the underlying model is trained twice and these models do not necessarily have to be the same. This results in additional variability to nonconformity measures, and thus this measure is not well suitable for TCP.

This measure was originally proposed by Papadopoulos and colleagues for ridge regression (Papadopoulos et al., 2002). It was used in combination with Support Vector Machine and Random Forests in Carlsson et al. (2014). With this normalized nonconformity measure, the length of prediction intervals will be proportional to the predicted accuracy of the underlying method at the new example. This means that we can make the conformal prediction more efficient, by providing larger prediction intervals for difficult examples and smaller ones for the easier examples to predict. The difficulty is defined as the accuracy of the underlying method. Using this locally weighted nonconformity measure, conformal prediction can produce locally-weighted prediction intervals, as shown in Lei et al. (2018). Focus on this fact, Bellotti (2020) combined this normalized nonconformity measure with a *surrogate* conformal predictor optimization, which is similar to ICP approximately though does not guarantee validity. Nevertheless, it was shown that the predictive efficiency for regression problems using several data while retaining validity was improved.

This normalized nonconformity measure was used in several applications, for example, change point detection (Ho and Wechsler, 2010), chemical engineering (Jablonka et al., 2020), deep learning (Cortés-Ciriano and Bender, 2019). Eklund et al. (2015) used this nonconformity measure Equation (4) for the drug discovery process and reported that ICP with the normalized nonconformity measure empirically shows more efficient prediction interval, even when *i.i.d.* assumption is often violated. However, there are no reports on direct comparisons between conformal predictors using different nonconformity measures. As such, it is not yet fully possible to assess the added value of this normalized nonconformity measure based on currently available literature.

In Papadopoulos and Haralambous (2011), a variant of normalized nonconformity measure was proposed as follows:

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\sigma_i + \beta} \tag{6}$$

where $\beta \geq 0$ works as a regularizer. With this normalized nonconformity the prediction interval for x_i is obtained as follows:

$$C(x_i) = (\hat{y}_i - \alpha_{s(\epsilon)}(\sigma_i + \beta),\ \hat{y}_i + \alpha_{s(\epsilon)}(\sigma_i + \beta)) \tag{7}$$

where $s(\epsilon) = \lfloor \epsilon(n+1) \rfloor$. By increasing β, we reduce the importance of σ_i and as a result, increases the importance of all other examples (Papadopoulos et al., 2007). In Papadopoulos

and Haralambous (2011), the effect of different values of β (0 and 0.5) on different datasets was investigated. The paper shows that the tighter prediction intervals can be obtained by adjusting β for each dataset. However, no hyperparameter tuning for β was performed. Johansson et al. (2014) compared the efficiency by tuning β and claimed that, although the parameter value is very important and dependent on the target range, all reasonable small values produce conformal predictors with similar efficiency.

Papadopoulos and colleagues proposed a total of six variants of normalized nonconformity measures in combination with k-nearest neighbors (Papadopoulos et al., 2011). For these measures, two quantities λ_i^k and ξ_i^k were calculated using k-nearest neighbors. For each example (x_i, y_i), T_i is the training data set which is used for predicting $\hat{y}$.

Firstly, λ_i^k measures expected accuracy based on the distance of the example from its k-nearest neighbours and was defined as:

$$d_i^k = \sum_{j=1}^{k} \delta(x_i, x_{ij}), \tag{8}$$

$$\lambda_i^k = \frac{d_i^k}{\text{median}(\{d_j^k : z_j \in T_i\})}, \tag{9}$$

where δ is a distance and d_i^k is the sum of the distance between x_i and its k-nearest neighbors, and is normalized with the distances of all training examples from its k-nearest neighbors.

Using λ_i^k, the following two nonconformity measures were defined as:

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\gamma + \lambda_i^k} \right|, \tag{10}$$

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\exp(\gamma \lambda_i^k)} \right|, \tag{11}$$

where γ works as a regularizer. While Equation (10) works in a similar way to the firstly introduced normalized nonconformity measure (Equation (6)), Equation (11) has a different property. This measure has a minimum value of 1 and the nonconformity score exponentially increases. As a result, this measure is more sensitive to changes when λ_i^k is large. The preferred measure depends on the context.

Secondly, ξ_i^k measures accuracy based on the standard deviation of the outputs of its k-neighbours and was defined as follows:

$$s_i^k = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \left(y_{ij} - \frac{1}{k} \sum_{j=1}^{k} y_{ij} \right)}, \tag{12}$$

$$\xi_i^k = \frac{s_i^k}{\text{median}(\{s_j^k : z_j \in T_i\})}. \tag{13}$$

Similarly, the following nonconformity measures were defined as below:

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\gamma + \xi_i^k} \right|, \tag{14}$$

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\exp(\gamma \xi_i^k)} \right|. \tag{15}$$

where again γ acts as a regularizer.

Finally, two quantities λ_i^k and ξ_i^k are combined and the following two nonconformity measures were proposed:

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\gamma + \xi_i^k + \lambda_i^k} \right|, \tag{16}$$

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\exp(\gamma \xi_i^k) + \exp(\rho \lambda_i^k)} \right|. \tag{17}$$

Parameters γ and ρ are used to control the sensitivity of ξ_i^k and λ_i^k, depending on the context. They showed that all the proposed nonconformity measures increased efficiency and especially two measures (Equation (16) and Equation (17)) are superior to the others.

Although these measures were developed for k-nearest neighbors, it is possible to apply the principal idea of these measures to different machine learning methods. Some nonconformity measures (Equation (16) and Equation (17)) have other parameters in addition to β. However, no tuning of these parameters was described either. Additionally, it is worth noting that the normalized noncofnormity measure is not widely adopted even when using the measure can potentially be helpful to provide efficient prediction intervals. One issue here might be that extending the normalized nonconformity measure is not straightforward in some application settings (e.g., high dimensional data for time series forcasting (Stankeviciute et al., 2021)).

More importantly, although the use of additional parameters in the normalized nonconformity measure has the potential to increase the efficiency (Boström et al., 2016), there is no clear indication of how this can be achieved. Therefore, at this moment, it is difficult to conclude which nonconformity measure behaves better in terms of the efficiency of the prediction intervals.

3.2. Quantile-based nonconformity measures

Although the normalized nonconformity measure can provide us with different sizes of predictive regions depending on a local estimate of variance, the following issues still remain.

Firstly, as mentioned by Lei and colleagues, the locally adaptive nonconformity measure cannot generate efficient prediction intervals when the data is homoscedastic (Lei et al., 2018). They argued that this inefficiency is mainly due to the extra variability arising from estimating σ_i (as can be seen from e.g., Equation (4)). Secondly, in general, experiments seem rather limited to draw solid conclusions about the advantage of using normalized nonconformity measures in many cases. Therefore, experiments using non-Gaussian distributions (e.g. multivariate, heavy-tailed and skewed) with different data sizes are required

to evaluate the performance of these normalized nonconformity measures. Knowledge obtained from such experiments can be used to determine whether normalized nonconformity measures will result in more efficient prediction intervals.

In order to overcome these issues discussed above, a new approach has been proposed that combines ICP with quantile regression (Romano et al., 2019). They mentioned that there is no guarantee that prediction intervals produced by quantile regression satisfy the desired coverage rate. This shortcoming motivated them to propose *Conformalized Quantile Regression* which requires a new quantile-based nonconformity measure.

During CQR, data are split into a training set L_1 and a calibration set L_2. Given any quantile regression algorithm M, lower and upper conditional quantile functions $\hat{q}_{\epsilon_{low}}$ and $\hat{q}_{\epsilon_{high}}$ are fitted to the proper training set:

$$\{\hat{q}_{\epsilon_{low}}, \hat{q}_{\epsilon_{high}}\} \leftarrow M(\{x_i, y_i\} : i \in L_1).$$

Using the calibration set, the conformity score is computed to quantify the error made by the plug-in prediction interval $\hat{C}(x) = [\hat{q}_{\epsilon_{low}}, \hat{q}_{\epsilon_{high}}]$ and are evaluated on the calibration set as follows:

$$\alpha_i = max\{\hat{q}_{\epsilon_{low}}(x_i) - y_i, y_i - \hat{q}_{\epsilon_{high}}(x_i)\}. \tag{18}$$

Finally, the prediction interval for x_i is constructed as follows:

$$C(x_i) = (\hat{q}_{\epsilon_{low}}(x_i) - Q_{1-\epsilon}(\alpha, L_2),\ \hat{q}_{\epsilon_{high}}(x_i) + Q_{1-\epsilon}(\alpha, L_2)). \tag{19}$$

where

$$Q_{1-\epsilon}(\alpha, L_2) \equiv (1-\epsilon)(1+1/|L_2|)\text{-th empirical quantile of } \{\alpha_i : i \in L_2\}. \tag{20}$$

The main difference between the previously introduced normalized nonconformity measures and quantile-based nonconformity measures is that these focus on estimating the lower and upper conditional quantiles rather than the conditional mean. The advantage of this approach is that it is fully adaptive to heteroscedasticity and theoretically guarantees a valid coverage. Moreover, their experiments showed more efficient prediction intervals compared to the normalized nonconformity measure as defined in Equation (4). However, to show that this method also works sufficiently efficient in the case of homoscedastic noise, further comparative research is needed. To the authors knowledge, such results have not been reported yet. Finally, it is also noteworthy that if quantile regression does not produce valid estimates, which can happen (Romano et al., 2019), the performance of CQR is also affected since it tries to cover the garuanteed validity by sacrificing efficiency (Chung et al., 2021).

A nonconformity measure similar to the quantile-based nonconformity measure (Romano et al., 2019) was proposed by Kivaranovic et al. (2020). Their measure was adapted to neural networks and to a wide class of data distributions. However, when using this quantile-based nonconformity measure, the validity of obtained conformal prediction intervals only holds under the assumption that the observations are *i.i.d.*. Such an assumption is not required for traditional conformal prediction methods and could limit the use of CQR in specific situations where such a condition cannot be met.

4. Influence of various factors on conformal prediction

Although the choice of nonconformity measure can considerably affect the validity and efficiency of prediction intervals, the exact influence often remains elusive. It is known that several factors (e.g., data, underlying model and nonconformity measure) can result in violation of the validity or maintain validity by sacrificing efficiency. Furthermore, some of these factors are entangled with each other, further complicating the choice of an appropriate nonconformity measure under different circumstances. Nevertheless, how these factors should be taken into account when choosing the nonconformity measure and influence the performance of conformal prediction has not been well explored yet.

This section attempts to illustrate this complexity of conformal prediction by highlighting these different influences and showing that it is not straightforward to choose a suitable nonconformity measure. In order to systematically evaluate these influencing factors, the effect on each step of the conformal prediction algorithm is discussed, as showing in Algorithm 1 (based on Sun, 2022). Although Algorithm 1 focuses on ICP, the same steps, except for splitting the whole dataset, are applicable to TCP. Based on Algorithm 1, we define two critical points where specific choices or circumstances can influence the performance of conformal prediction namely: the choice of nonconformity measure and how data is split into proper training set and calibration set.

Algorithm 1: Inductive Conformal Prediction

Input: Dataset $D = (x_i, y_i)_{i=1}^{n}$ (Section 4.2), underlying model M, nonconformity measure α (Section 4.1), test point x_{n+1}, target confidence level $1 - \epsilon$

Output: Prediction interval $\Gamma^{1-\epsilon}(x_{n+1})$

Randomly split dataset D into training and calibration datasets $D = D_{train} \cup D_{cal}$ where $|D_{train}| = m$ and $|D_{cal}| = n - m$;

Train underlying model $\hat{f} = M(D_{train})$;

Initialize nonconformity scores $\alpha_{cal} = \{\}$;

for $(x_i, y_i) \in D_{cal}$ **do**

  $\alpha_{cal} \leftarrow s \cup \{S((x_i, y_i), \hat{f})\}$

end

Calculate the $(1 - \epsilon)-$th quantile $q_{1-\epsilon}$ of $\alpha_{cal} \cup \{\infty\}$

return $\Gamma^{(1-\epsilon)}(x_{n+1}) = \{y : S((x_{n+1}, y), \hat{f}) \leq q_{1-\epsilon}\}$

4.1. Choosing a nonconformity measure

Nonconformity measures can be grouped into *model-agonistic* and *model-dependent* (Aleksandrova and Chertov, 2021b). However, Papadopoulos et al. (2002) mentioned that, in many situations, the nonconformity measure is chosen based on the underlying algorithm and thus it is natural to expect that the more accurate underlying models lead to better conformal predictions. More specifically, this step influences the performance of conformal prediction regarding efficiency of prediction intervals, as shown by Romano et al. (2019). This was also confirmed by Aleksandrova and Chertov (2021b,a) for a classification setting. In addition, Laxhammar and colleagues suggest that the use of domain knowledge (e.g., knowledge about the data distribution) is of importance when defining nonconformity

measures for specific applications (Laxhammar and Falkman, 2010). Therefore, one should carefully consider the choice of nonconformity measure since it depends strongly on the context of the problem (Vovk et al., 2005). However, systematic evaluation of different nonconformity measures under different conditions or context has not yet been performed.

4.2. How to split the data

In ICP, the whole dataset has to be divided into a proper training dataset and a calibration set. Angelopoulos and colleagues confirmed that size of the calibration dataset has an influence on the validity of conformal prediction (Angelopoulos and Bates, 2021). Their experiments showed that larger data size leads to more stable prediction intervals, and thus improved validity. With increasing size of the calibration set, it is possible to choose nonconformity measures with additional hyperparameters, with the possibility to improve efficiency of the prediction intervals. These hyperparameters can in turn be tuned using the calibration set to improve the overall performance of the conformal prediction using this specific nonconformity measure. This means that the size of the calibration set has an impact on the choice of nonconformity measures.

However, increasing the size of the calibration set at the expense of the proper training set size negatively influences the choice and training of the underlying model. With the knowledge that some models require significantly more training, reducing the proper training set would limit the choice of models that can be used. In order to overcome this issue, out-of-bag conformal methods, where out-of-estimates are used for the calibration set, have been introduced (e.g., Johansson et al. (2014); Gupta et al. (2022)). Although this approach could improve the training process of the model, the risk is that optimizing hyperparameters of nonconformity measure on these alternative calibration sets lead to violations of the exchangeability assumption.

It remains difficult to quantitatively assess the impact of proper training and calibration set sizes on a choice of nonconformity measures and the resulting validity and efficiency of the prediction intervals under different circumstances. This warrants more detailed experiments to understand the influence of the size of calibration and proper training set.

5. Discussion and future direction

The advantage of conformal prediction (i.e. distribution-free) makes it suitable to apply to different (scientific) problems since it does not require assumptions on output distributions, which is rarely given as a-prior. Indeed, several authors show the use of conformal prediction in a variety of different applications. However, this will also implicitly result in the requirement that nonconformity measures should deal with data having different characteristics (e.g. size, dimensionality, noise distribution). In order to accustom these requirements, new nonconformity measures have been developed. Many of these (recent) nonconformity measures focus on a specific characteristic of the dataset in question. In this section, we discuss on some of the challenges associated with nonconformity measures to handle specific datasets and suggest possible future directions for the development of new nonconformity measures.

5.1. Challenges faced in choosing an appropriate nonconformity measure

As described in Section 3, non-normalized nonconformity measures tend to lose efficiency of the prediction intervals when data has heteroscedastic noise. In order to overcome this issue, normalized nonconformity measures and quantile-based nonconformity measures were developed. The additional parameters (e.g., β in Equation (6)) are used to adjust the prediction intervals to individual datasets and experiments. Although we can obtain tighter prediction intervals using such nonconformity measures, increasing the number of parameters may result in increased interval variability. This means that there is a trade-off between validity and efficiency with respect to the additional parameters. Therefore, it is important to investigate the effect of including additional parameters on the performance of conformal prediction in terms of this validity-efficiency trade-off. Furthermore, it is interesting to investigate whether it is possible to draw a parallel between this validity-efficiency trade-off and the well-known bias-variance trade-off. In other words, whether the validity-efficiency trade-off can be related to or interpreted as the bias-variance trade-off is still in question and requires further investigation.

While the validity of prediction intervals using normalized nonconformity measures and quantile-based measures, is empirically proven in the literature, a more detailed theoretical understanding is required. In particular, knowledge about the performance of nonconformity measures, where parameters are to be extensively tuned on the training data, is currently lacking. In addition, there is also a lot to be gained in an empirical setting. In current literature there have been limited in-depth assessments of the advantages and disadvantages of the various nonconformity measures. However, we believe that this is an important step towards a broader adoption of conformal prediction across different fields.

Lastly, during the process of both choosing and developing nonconformity measures, it is important to remind that we cannot hope for one nonconformity measure to work with all problems. In this regard, we suggest that the choice of nonconformity measure – or development of a new nonconformity measure for that matter – has to be done mainly based on data type. Choosing an optimal nonconformity measure can then be done by utilizing information from domain knowledge or previous experiments.

5.2. Future development of new nonconformity measure

One direction where an alternative nonconformity measure can be useful, is in the case of data containing non-Gaussian heteroscedastic noise, or outliers. In this regard, conformal prediction with a quantile-based nonconformity measure (Section 3.2) has been proposed (Romano et al., 2019; Kivaranovic et al., 2020). Although the quantile-based nonconformity measure has the potential to deal with a different types of data noise or outliers, this measure requires output from quantile regression method. Therefore, in case of limited amount of data, there is a possibility that the performance of quantile regression suffers too much from outliers so that the efficiency of prediction intervals is sacrificed unnecessary. Additionally, without a sufficient amount of data, the validity of the prediction intervals from quantile regression is not guaranteed. In this case, it is possible that conformal prediction with quantile-based nonconformity measures, using quantile regression as the underlying algorithm, suffers from inefficient prediction intervals in order to retain the validity. This

highlights a general issue in conformal prediction regarding the validity-efficiency trade-off, which is also discussed in Section 5.1.

Regarding the aforementioned trade-off, it could be beneficial to design a nonconformity measure that allows one to explicitly control the balance between validity and efficiency. Potentially, this can be achieved by a tunable nonconformity measure that considers validity and efficiency separately. It may then be possible to the balance these two quantities depending on the context. This idea is inspired by Chung et al. (2021), where new quantile methods were proposed for calibrated uncertainty quantification. Their underlying motivation is that validity should be first achieved and then efficiency optimized. This is exactly what we would like to achieve in conformal prediction using a new nonconformity measure. Their results showed that the proposed methods provide better means of learning calibrated conditional quantiles.

They focus on *pinball loss*. Given a target y, a prediction $\hat{y}$ and quantile level $\tau \in (0, 1)$, the pinball loss ρ_τ is defined as

$$\rho_\tau(y, \hat{y}) = (\hat{y} - y)(\mathbf{1}\{y \leq \hat{y}\} - \tau).$$

Although many current quantile-based methods focus on optimizing the pinball loss, they highlight some limitations of the pinball loss. They claimed that, although the pinball loss targets both validity and efficiency, the balance of these two quantities is made implicitly, resulting in a poor optimization objective. In order to mitigate the shortcoming, they design several quantile methods that consider two objectives (validity and efficiency) separately, then these two objectives are combined into a single loss function. This loss function can provide an explicit balance between validity and efficiency that can be chosen by the end user. We believe that we can apply this idea to nonconformity measures in order to balance validity and efficiency.

Lastly, it is worth mentioning that some studies focus on the trade-off between validity and efficiency using existing nonconformity measures. For instance, Lei and Bellotti (2023) used directly optimized inductive conformal regression, using Equation (3) as nonconformity measure, which takes only the average width of prediction intervals as the loss function and increases the efficiency while retaining validity.

6. Conclusion

We discussed the use of nonconformity measures for conformal prediction in a regression setting. Given that the nonconformity measure is one of the important ingredients for conformal prediction, there have only been relatively few different nonconformity measures developed. Therefore, we argue that it is time to shift our attention to nonconformity measures, more specifically the development and study of nonconformity measures under different circumstances. Future research is required to establish the relation between nonconformity measures and the resulting performance of the conformal prediction.

Although there are still many unknowns regarding the accuracy and performance of conformal prediction when using different nonconformity measures, there are promising results in the literature showing the use of conformal prediction in different practical situations. We believe that conformal prediction increasingly play a more important role in uncertainty quantification in future.

References

Marharyta Aleksandrova and Oleg Chertov. How nonconformity functions and difficulty of datasets impact the efficiency of conformal classifiers. August 2021a.

Marharyta Aleksandrova and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. 152:151–170, 2021b.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and Distribution-Free uncertainty quantification. *arXiv [cs.LG]*, July 2021.

Anthony Bellotti. Constructing normalized nonconformity measures based on maximizing predictive efficiency. 128:41–54, 2020.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Evaluation of a Variance-Based nonconformity measure for regression forests. In *Conformal and Probabilistic Prediction with Applications*, pages 75–89. Springer International Publishing, 2016.

Lars Carlsson, Martin Eklund, and Ulf Norinder. Aggregated conformal prediction. In *Artificial Intelligence Applications and Innovations*, pages 231–240. Springer Berlin Heidelberg, 2014.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. October 2021.

Isidro Cortés-Ciriano and Andreas Bender. Deep confidence: A computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.*, 59(3):1269–1281, March 2019.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.*, 74(1-2):117–132, June 2015.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *BJOG*, 29(1):1–23, February 2023.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognit.*, 127:108496, July 2022.

Shen-Shyang Ho and Harry Wechsler. A martingale framework for detecting changes in data streams by testing exchangeability. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32 (12):2113–2127, December 2010.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-Data science in porous materials: Materials genomics and machine learning. *Chem. Rev.*, 120(16):8066–8129, August 2020.

Ulf Johansson, Rikard König, Tuve Löfström, and Henrik Boström. Evolved decision trees as conformal predictors. In *2013 IEEE Congress on Evolutionary Computation*, pages 1794–1801, June 2013.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Mach. Learn.*, 97(1):155–176, October 2014.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv [cs.LG]*, September 2022.

Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. Adaptive, Distribution-Free prediction intervals for deep networks. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4346–4356. PMLR, 2020.

Rikard Laxhammar and Göran Falkman. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, StreamKDD '10, pages 47–55, New York, NY, USA, July 2010. Association for Computing Machinery.

Haocheng Lei and Anthony Bellotti. Reliable prediction intervals with directly optimized inductive conformal regression for deep learning. February 2023.

Jing Lei, James Robins, and Larry Wasserman. Distribution free prediction sets. *J. Am. Stat. Assoc.*, 108(501):278–287, 2013.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.*, 74(1-2):29–43, June 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-Free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523):1094–1111, July 2018.

Henrik Linusson. *Nonconformity Measures and Ensemble Strategies : An Analysis of Conformal Predictor Efficiency and Validity*. PhD thesis, 2021.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *Artificial Intelligence Applications and Innovations*, pages 261–270. Springer Berlin Heidelberg, 2014.

H Papadopoulos, V Vovk, and A Gammerman. Regression conformal prediction with nearest neighbours. *jair*, 40:815–840, April 2011.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. InTech, August 2008.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Netw.*, 24(8):842–851, October 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer Berlin Heidelberg, 2002.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, volume 2, pages 388–395, October 2007.

Yaniv Romano, Evan Patterson, and Emmanuel J Candès. Conformalized quantile regression. *arXiv [stat.ME]*, May 2019.

C Saunders, A Gammerman, and V Vovk. Transduction with confidence and credibility. pages 722–726. eprints.soton.ac.uk, 1999.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *arXiv [cs.LG]*, pages 371–421, June 2007.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Adv. Neural Inf. Process. Syst.*, 34:6216–6228, December 2021.

Sophia Sun. Conformal methods for quantifying uncertainty in spatiotemporal data: A survey. September 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *arXiv [stat.ME]*, April 2019.

V Vovk, A Gammerman, and G Shafer. Conformal prediction. In Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, editors, *Algorithmic Learning in a Random World*, pages 17–51. Springer US, Boston, MA, 2005.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C H Hoi and Wray Buntine, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore, 2012. PMLR.

Vladimir Vovk. Cross-conformal predictors. *Ann. Math. Artif. Intell.*, 74(1-2):9–28, June 2015.

Vladimir Vovk, Ilia Nouretdinov, Valentina Fedorova, Ivan Petej, and Alex Gammerman. Criteria of efficiency for conformal prediction. March 2016.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Modifications of conformal predictors. In Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, editors, *Algorithmic Learning in a Random World*, pages 107–142. Springer International Publishing, Cham, 2022.