Proceedings of Machine Learning Research 230:1–15, 2024 Conformal and Probabilistic Prediction with Applications

Conformal Stability Measure for Feature Selection Algorithms

Marcos López-De-Castro* MLOPEZDECAS@UNAV.ES
Alberto García-Galindo AGARCIAGALI@UNAV.ES
Rubén Armañanzas RARMANANZAS@UNAV.ES
Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Ismael Sánchez Bella Building, Campus Universitario, 31009 Pamplona, Spain.
TECNUN School of Engineering, Universidad de Navarra, Donostia-San Sebastián, Spain.

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Quantifying the stability of feature selection techniques has been an ongoing challenge over the last two decades. A large number of stability estimators have been proposed to overcome this problem, but performance guarantees based on suitable statistical frameworks are lacking. A recently developed framework proposed a new and robust estimator of the stability and a method to quantify the uncertainty of the estimates through approximate confidence intervals. Unfortunately, this statistical framework is based on asymptotic assumptions. In situations in which a low number of subsets of selected features are available for the quantification of the stability estimator, the coverage guarantees provided by this framework do not hold. In this work, we propose a method to estimate stability and achieve validity in a situation where only a few samples are available. We take advantage of the Conformal Prediction framework, constructing prediction intervals without any assumption about the underlying distribution of data. Extensive simulations show that our method successfully achieves conservative validity. Furthermore, as the number of available samples increases efficiency is also achieved. Comparisons between prediction intervals and confidence intervals show an acceptable trade-off between coverage guarantees and the interval length for the former, while there is a clear miscoverage for the latter.

Keywords: Conformal prediction · Feature selection · Stability measure · Coverage guarantees

1. Introduction

Feature selection is a well-known approach to overcome the curse of dimensionality. This technique, which is widely used by data scientists and machine learning engineers, reduce the dimensionality of data without altering the original representation of features (Saeys et al., 2007). This property allows us to identify meaningful scientific knowledge, remove irrelevant features to improve prediction performance, or decrease computational performance eliminating redundant features. A broadly used classification of these methods is based on its evaluation strategy: filter, wrapper, and embedded methods, although mixed approaches have also been proposed (Guyon et al., 2006; Bielza and Larrañaga, 2020). Filter methods evaluate the intrinsic properties of data to find correlations and statistical

* Corresponding author

dependences, and are independent from the classification methodology. An example of representative methods from this family are those based on mutual information (Brown et al., 2012). Wrapper techniques use the performance of a classifier to find relevant features. This family provides competitive results usually at high computational cost. Moreover, features selected are dependent on the classifier's performance criteria. Lastly, in embedded methods, the selection of features is guided by the training process of a prediction model. This reduces the computational cost with respect to wrappers. Well-known methods belonging to this family are the LASSO regularization (Tibshirani, 1996) and the tree-based methods Breiman (2001).

The evaluation of feature selection results typically involves assessing the improvement of a predefined performance score of a learning algorithm, *e.g.*, accuracy. However, if the selection method is applied to another subsample of the dataset, the question arises as to how stable the feature selection method is. Stability quantifies the sensitivity of a method to small changes in the training set. The stability of feature selection algorithms is crucial, particularly in biomedical domains (Davis et al., 2006; Jurman et al., 2008). This is because the final goal of these applications is usually to find a small set of highly discriminatory features for later exploration (Kalousis et al., 2007). If a significant change in the selected features is observed with only minor changes in the training dataset, the algorithm's findings may not be robust. It is worth noting that in high-dimensional and complex problems, such as predicting phenotype from molecular signatures, the phenomenon of multiplicity may appear (Statnikov and Aliferis, 2010; Statnikov et al., 2013). This happens whenever two or more completely different subsets of non-redundant features are found to be maximally predictive of a phenotype. The causes behind this phenomenon are not clear, but throughout this work, we assume that there is only one subset of non-redundant features that maximizes the performance of a given predictor.

Stability has been extensively studied in the field of learning algorithms. However, the pioneering, in-depth work on stability for feature selection algorithms was carried out by Kalousis et al. (2005, 2007). Since these studies, a large number of stability measures have been proposed in the literature (see Nogueira (2018) for more in-depth discussion). Among all the proposals, one that should be highlighted is the stability measure put forward by Kuncheva (2007), which has become a standard. In the present study, we focus on the stability measure introduced by Nogueira et al. (2018), which generalizes Kuncheva's measure. This new stability measure allows the quantification of the stability between subsets of features with different cardinalities.

Contribution of this work: To ensure a fair stability comparison between two or more feature selection algorithms, the ability to quantify the uncertainty of the estimator is crucial. Coverage guarantees become imperative at this point. Nogueira's proposed stability estimator can be associated with a sampling distribution. This allows the derivation of approximate confidence intervals for stability estimates. However, in their statistical framework, the underlying distribution is only approached asymptotically. Unfortunately, a common situation in feature selection is the limited availability number of selected subsets. Factors that contribute to this limitation include: a large amount of data, the computational

complexity of the selection algorithm, or the requirement for an exhaustive search. Usually, no more than 5 to 10 independent subsets of selected features are available, *e.g.*, in a cross-validation procedure. In this work, we focus on this common scenario where the number of available samples is insufficient to satisfy assumptions based on the multivariate central limit theorem. We propose using the conformal prediction framework (Vovk et al., 2005; Shafer and Vovk, 2008) to provide efficient, valid, and finite-sample prediction intervals to quantify the uncertainty surrounding the stability estimate.

Outline: The rest of the manuscript is organized as follows. Section 2 introduces Nogueira's and Conformal Prediction frameworks. In Section 3, our methodological proposal is presented. Section 4 describes the experimental setup, while the results are detailed and discussed in Section 5. Finally, Section 6 concludes the paper and presents future directions of research.

2. Preliminaries

2.1. Nogueira's stability estimator

In their original work, Nogueira et al. (2018) defined a set of five properties that a good stability estimator must have. They found that none of the stability estimators previously proposed in the literature had all of these properties, and defined a new estimator that satisfied them, namely, fully defined, strict monotonicity, known bounds, maximum stability if - and only if - the selection is deterministic, and correction for chance. Let's assume we are given a dataset $\mathcal{D}$ for a classification task $\mathcal{D} = \{(X, Y)\}$, where $X \in \mathbb{R}^d$ is the set of covariables and Y the prediction target. Throughout this work, we denote a specific feature in the covariate space as the random variable $X^j,\ \forall j \in \{1, ..., d\}$. Let $\pi(\cdot)$ be a feature selection method so that $\pi(D) = z$, where z is a binary string of length d. A value of 1 in the j^{th} position means that the feature X^j has been selected, whereas a 0 means that it has not. We take M bootstrap samples from D and apply the feature selection method π to each bootstrap sample, obtaining a collection of feature sets $\mathcal{Z} = \{z_1, ..., z_M\}$. The collection of feature sets $\mathcal{Z}$ can be thought of as a matrix of size $M \times d$. Each realization of the feature selection method produces a set of features regardless of any previous realization of the method, so it is plausible to assume independence, in the sense of no having data leakage, between the rows of matrix $\mathcal{Z}$. From this assumption, we can infer the true stability Φ of a feature selection method when the columns of matrix $\mathcal{Z}$ are modeled as random variables drawn from a Bernouilli distribution with mean parameters b_j:

$$\Phi = 1 - \frac{\frac{1}{d}\sum_{j=1}^{d} b_j(1 - b_j)}{\bar{b}(1 - \bar{b})}, \tag{1}$$

where $\bar{b} = \frac{1}{d}\sum_{j=1}^{d} b_j$, and b_j models the probability of a feature X^j to be selected by π. The mean parameters b_j are usually not known in real world applications, so stability must be estimated.

Definition 1 (Stability estimator) *A stability estimator for feature selection algorithms is as follows:*

$$\hat{\Phi}_N(\mathcal{Z}) = 1 - \frac{\frac{1}{d}\sum_{j=1}^{d} s_j^2}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}, \tag{2}$$

where $s_j^2 = \frac{M}{M-1}\hat{b}_j(1-\hat{b}_j)$, $\hat{b}_j = \frac{1}{M}\sum_{i=1}^{M} z_{ij}$, $\bar{k} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{d} z_{ij}$ *and* z_{ij} *is the element* $i,j)$ *of the matrix* $\mathcal{Z}$.

In Nogueira et al. (2018), it was shown that this stability estimator satisfies the five following properties required to be a good stability estimator:

1. **Fully defined:** The estimator can deal with feature selection methods that return a variate number of features.

2. **Strict monotonicity:** The estimator is an increasing function of the size of the average pairwise intersection size $\frac{1}{M(M-1)}\sum_{l}^{M}\sum_{k\neq l}^{M}|z_l \cap z_k|$ between two rows z_l, z_k.

3. **Bounded:** The stability score produced by the estimator is bounded. This is required for a meaningful interpretation and enables the stability comparison. The estimator defined in Expression 2 is bounded within $[-\frac{1}{M-1}, 1]$.

4. **Maximum stability $\iff$ Deterministic selection:** When two sets of selected features are identical, their similarity is maximal (Kuncheva, 2007). This implies that the estimator must reach its maximum score if, and only if, all the selected feature sets are identical in size and elements.

5. **Correction for chance:** The expectation of the stability estimator remains constant whenever we have independently drawn subsets at random. This property is known as *correction for chance*, and reflects on the similarity of feature subsets that occurred by chance.

In addition, they showed that their estimator has the same statistical properties as Fleiss' kappa (Fleiss, 1971). Following the work of Gwet (2008), they proved that confidence intervals can be asymptotically ($M \to \infty$) derived because the statistic $\hat{\Phi}_N$ weakly converges to a normal distribution.

Definition 2 ($\hat{\Phi}_N$ confidence interval) *A* $(1-\alpha)$*-approximate confidence interval for* $\hat{\Phi}_N$ *is*

$$[\hat{\Phi} - z^*_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\sigma_{\hat{\Phi}}}\ ,\ \hat{\Phi} + z^*_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\sigma_{\hat{\Phi}}}], \tag{3}$$

where $z^*_{\left(1-\frac{\alpha}{2}\right)}$ *is the inverse cumulative of a standard normal distribution at* $1-\frac{\alpha}{2}$ *and* $\sqrt{\sigma_{\hat{\Phi}}}$ *is an estimate of the variance.*

Further details, including the demonstrations, are available at Nogueira et al. (2018) and Nogueira (2018).

2.2. Conformal prediction

Conformal prediction is an uncertainty quantification framework originally proposed by Vovk, Gamerman, Shafer and Vapnik (Gammerman et al., 1998; Vovk et al., 2005). Conformal prediction quantifies the uncertainty of a prediction providing valid and asymptotically efficient prediction intervals -or prediction sets in classification tasks-, instead of traditional point predictions (Shafer and Vovk, 2008; Balasubramanian et al., 2014; Angelopoulos and Bates, 2023). The framework provides finite-sample coverage guarantees,

i.e., given a dataset $\mathcal{D} = \{Z_i := (X_i, Y_i)\}_{i=1}^n$, an unknown new sample X_{n+1} and a confidence level $1-\alpha$, the property

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha, \tag{4}$$

holds (Toccaceli, 2022), where $\mathcal{C}(\cdot)$ denotes the prediction interval for X_{n+1}. This property is known as the *marginal validity*, and is achieved without making any assumptions about the underlying distribution of the samples. The only requirement is that these samples must be independent and identically distributed (i.i.d.) observations[1]. However, the length of the prediction intervals will depend on the choices we make when building the conformal algorithm. We want the prediction interval to be as narrow as possible in order to be informative, a property known as *efficiency*.

To illustrate how this framework performs, suppose that we are given a bag[2] of i.i.d. samples $\{Z_1, ..., Z_n\}$ from an unknown distribution $\mathcal{P}$. We want to construct a prediction interval for a new unknown sample $Z_{n+1} \sim \mathcal{P}$ with a confidence level of $1-\alpha$. Let's make the following hypothesis $H_0 : Z_{n+1} = z := (X_{n+1}, y)$, where y is some hypothetical value. At this point we need to define a central notion in conformal prediction: the non-conformity measure. A non-conformity measure is a real-valued function $A : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ that quantifies how strange a sample $Z_i \in \{Z_1, ..., Z_n, z\}$ is in the bag $\{Z_1, ..., Z_n, z\}$, *i.e.*, the degree to which the sample does not conform with the bag. A non-conformity measure assigns a numerical non-conformity score

$$\varphi_{i,z} = A(\{Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_n, z\}, Z_i), \tag{5}$$

to each sample $Z_i \in \{Z_1, ..., Z_n, z\}$. Non-conformity measures are typically defined as the residual of a fitted estimator $\hat{\theta}$ within the augmented set of samples $\{Z_1, ..., Z_n, z\}$. However, a non-conformity score φ_i alone does not inform us how strange the sample Z_i is. We need to compare φ_i with the other $\varphi_{j \neq i}$ non-conformity scores. Returning to our main point of interest, *i.e.*, establishing which are the possibilities that our trial value z could be Z_{n+1}, we can determine how strange z is if we count the number of samples in the bag that have a large non-conformity score. Note that if we normalize this number with respect to the number of elements in the bag,

$$p^z = \frac{|\{i = 1, ..., n+1 \mid \varphi_i \geq \varphi_{n+1}\}|}{n+1}, \tag{6}$$

we can conveniently define a p-value for our hypothesis test. This is a valid p-value because the i.i.d. property is required on the samples in the bag $\{Z_1, ..., Z_{n+1}\}$, forcing the vector of non-conformity measures to be exchangeable. This implies that if H_0 is true, the vector of non-conformity measures is uniformly distributed among $\{1, ..., n+1\}$, so that p^z is also uniformly distributed over $\{\frac{1}{n+1}, ..., 1\}$. For this reason, p^z is a valid p-value in the sense that $\mathbb{P}(p^z \leq \alpha) \geq 1 - \alpha$. Finally, if we wish to build the prediction interval $\mathcal{C}(X_{n+1})$ with a

1. The only hard requirement is that data must be exchangeable, but the i.i.d. assumption implies exchangeability and is a common assumption, specifically among machine learning practitioners.
2. A bag is a set that allows for repeated elements.

confidence level of $1-\alpha$, we only need to invert the hypothesis test and accept all the trial values z that meet:

$$\mathcal{C}(X_{n+1}) = \{y : p^z \leq \alpha\}. \tag{7}$$

It should be noted that in classification problems, the set of trial values is all the possible classes available, whereas in regression we need to define an appropriate interval of trial samples. Due to the above-mentioned factors this methodology, known as *transductive* conformal prediction, can be computationally demanding. Fortunately, a computationally competitive variant called *split* or *inductive* conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005) was developed with the same coverage guarantees, but sacrificing efficiency. In this work, we focus on transductive conformal prediction for regression problems.

3. Methodology

Let us suppose that we are given an $M \times d$ binary matrix $\mathcal{Z}$ as defined in Section 2.1. We are interested in estimating the stability of the feature selection method that produced $\mathcal{Z}$ using the estimator defined in (2). Moreover, we also want to obtain a valid estimate of the stability as defined in Equation (4), regardless of the number of rows in $\mathcal{Z}$. This is because the number of rows in $\mathcal{Z}$ reflects the number of available samples to estimate stability. We propose to take advantage of the conformal prediction methodology to achieve empirical coverage guarantee on this stability estimation problem.

Conformal prediction uses past experience to construct valid prediction intervals, so we need to find a way to discover this "past experience" from the given matrix $\mathcal{Z}$. As stated in Section 2.1, rows *i.e.*, samples, in the matrix $\mathcal{Z}$ are assumed to be independent. Nogueira's model assumes that the elements of each column in $\mathcal{Z}$ follow a Bernouilli distribution with parameter b_j. These parameters, although unknown, are fixed for each column as long as there is no multiplicity. Let us perform a subsampling of the matrix $\mathcal{Z}$ by rows, so that a dataset $\mathcal{R} = \{\mathcal{Z}_1, ..., \mathcal{Z}_c\}$ is generated, where each $\mathcal{Z}_i$ is a $\kappa \times d$ binary matrix with $\kappa < M$. Due to the independence of each row in $\mathcal{Z}$ and the fact that all the elements in a column j follow the same distribution, we can assume the subsamples to be indistinguishable.

Therefore, the stabilities of any two different subsamples, $\mathcal{R}_1$ and $\mathcal{R}_2$ must be equally distributed. So, we can split the original binary matrix into a set of matrices. Specifically, we compute the $\binom{M}{\kappa} = c$ different combinations of κ rows from $\mathcal{Z}$. Note that in Nogueira's framework, this procedure cannot be used to increase the available sample size for computing confidence intervals. This is due to the asymptotic validity of the confidence interval as $M \to \infty$, hence, validity will only improve if $\mathcal{Z}$ has larger M values.

Once we have identified a bag of samples drawn from the same underlying distribution, we are able to predict a valid prediction interval to quantify the uncertainty of any point estimate $\hat{\theta}_z$ of the random variable Φ based on the set of samples $\{\hat{\Phi}_N(\mathcal{Z}_1)), .., \hat{\Phi}_N(\mathcal{Z}_{i-1})), \hat{\Phi}_N(\mathcal{Z}_{i+1})), ..., \hat{\Phi}_N(\mathcal{Z}_n)), \hat{\Phi}_N(z))\}$. Prediction intervals are derived using the transductive conformal prediction approach described in Section 2.2. The next step is build the calibration samples. In order to maximize the number of calibration samples and preserve the

underlying distribution, we only take combinations of κ elements from $\mathcal{Z}$ for which c is maximum. If there are two κ for which c is maximal, we will choose the largest κ. Finally, we define the non-conformity measures (5) as

$$\varphi_{z,i} = f(\hat{\theta}_z, \hat{\Phi}_N(\mathcal{Z}_i)) \qquad \forall i \in \{1, ..., c\}, \tag{8}$$

$$\varphi_{z,c+1} = f(\hat{\theta}_z, \hat{\Phi}_N(z)), \tag{9}$$

where $\hat{\Phi}_N(\cdot)$ was defined in Expression (2), $\hat{\Phi}_N(z)$ is a proposed trial value in the interval $(-\frac{1}{\kappa-1}, 1)$ and f is a function that quantifies the distance between the point estimate and a sample. The described procedure is summarized in Algorithm 1.

Algorithm 1 Prediction of stability using the transductive approach.

Input: $\mathcal{D} = (\hat{\Phi}_N(\mathcal{Z}_1), ..., \hat{\Phi}_N(\mathcal{Z}_c))$;
Input: Trial values $\mathcal{Z}_{trial} = \{-\frac{1}{\kappa-1}, ..., 1\}$;
Input: significance level α;

for $\hat{\Phi}_N(z)_j \in \mathcal{Z}_{trial}$:

for $\mathcal{Z}_i \in \mathcal{D}$:

$\hat{\theta}_j \leftarrow \hat{\theta}(\{\mathcal{D} \cup \{\hat{\Phi}_N(z)_j\})/\{(\mathcal{Z}_i\})$ # compute point estimation

$\varphi_{j,i} \leftarrow f(\hat{\theta}_j, \hat{\Phi}_N(\mathcal{Z}_i))$ # compute the non-conformity scores

$\varphi_{j,c+1} \leftarrow f(\hat{\theta}_j, \hat{\Phi}_N(z)_j)$ # compute the non-conformity scores of the trial value

$p^j \leftarrow \dfrac{|\{i = 1, ..., c+1 \mid \varphi_{j,i} \geq \varphi_{j,c+1}\}|}{c+1}$ # compute and save the p-value

$\mathcal{C} \leftarrow \{\hat{\Phi}_N(z)_j \in \mathcal{Z}_{trial} : p^j > \alpha\}$ # compute prediction intervals with confidence $1 - \alpha$

Return: $\mathcal{C}$, *the valid prediction interval* for $\mu_{\hat{\Phi}_N(z)_j}$.

4. Experimental settings

Focusing on a scenario where only small samples are available and the coverage guarantees for the confidence interval defined in Expression (3) are broken, let's suppose a $M \times 100$ binary matrix $\mathcal{Z}$ with $M = m, \ \forall m \in \{5, ..., 10\}$[3]. We chose to study the cases from $M = 5$ up to $M = 10$ because this is a plausible situation when research based on computationally

3. The number of columns is irrelevant. We decided upon 100 columns in line with the work of Nogueira et al. (2018).

demanding feature selection algorithms is conducted. This may be, for example, the resulting matrix derived from applying a feature selection method on a cross-validation scenario. The tests are carried out on artificial datasets codified as the matrix $\mathcal{Z}$, whose columns, *i.e.*, the features, are drawn from a Bernoulli distribution with a known mean parameter p_j. This assures a known value for the true stability function Φ, which will be referred to as the *oracle*.

We now turn to the following two aspects: (i) the consistence of the estimations $\hat{\Phi}_N(\mathcal{Z}_i)$ for low values of κ, and (ii) the estimation of the stability from $\{\mathcal{Z}_1, ..., \mathcal{Z}_c\}$. In Figure 1, we show some examples of synthetic datasets for different combinations of M and κ. The stability estimate for each $\mathcal{Z}_i$ is shown in blue, while the oracle and the mean of the vector of estimates $(\hat{\Phi}(\mathcal{Z}_1), ..., \hat{\Phi}(\mathcal{Z}_c))$ are indicated in red and black, respectively. As we can see, the $\hat{\Phi}(\mathcal{Z}_i)$ values are close to the true stability predicted by the oracle $\Phi(\mathcal{Z})$, and the greater the κ the closer it is to the oracle. This last statement was also empirically observed by Nogueira et al. (2018). The second observation is about the mean of the stability's estimates. It seems very close to the true stability, so it is postulated as a simple but acceptable estimator candidate to provide point predictions. To strengthen these findings, we show in Figure 2 the residuals for 1000 independent samples using the combinations in Figure 1. The residuals are small, as shown in Figure 2. We observe that as the number of rows in the matrix $\mathcal{Z}$ increases, the residuals improve. This is because a larger number of samples leads to a more accurate estimate. Moreover, the number κ also has an impact on the quality of the estimations.

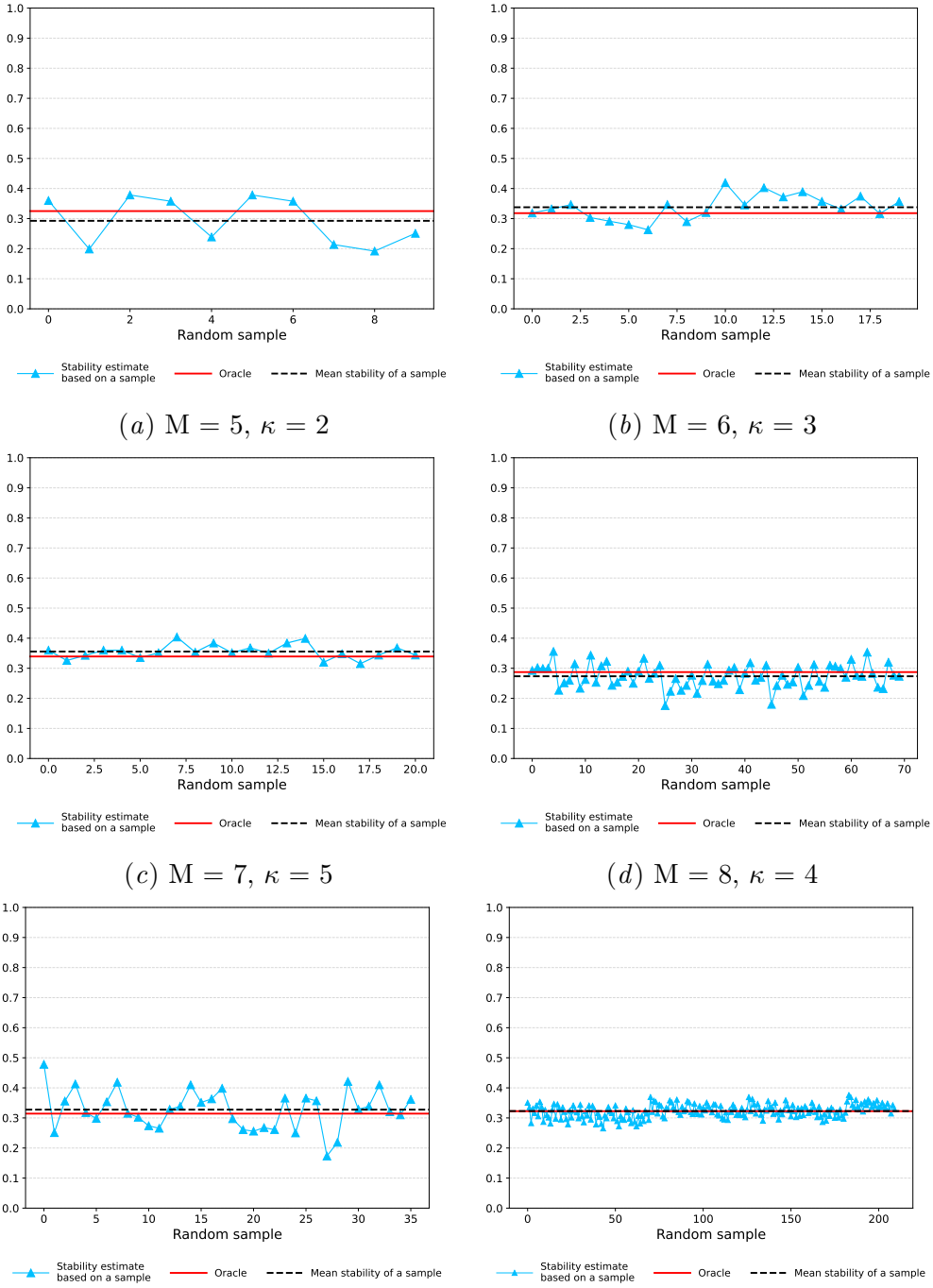
As suggested in the previous paragraph, we take the mean as point estimator. Finally, we define the following non-conformity measure

$$\varphi_{z,i} = |\frac{\hat{\Phi}(\mathcal{Z}_i) - \mu_z}{\sigma_z}|, \tag{10}$$

where μ_z, σ_z are the mean and the standard deviation of $\{\mathcal{Z}_1, ..., \mathcal{Z}_c, z\}$ and z is a trial value in the interval $(-\frac{1}{\kappa-1}, 1)$. We performed 1000 independent simulations for each m. On each simulation, we used 500 test values equally-spaced along the interval $(\frac{-1}{\kappa-1}, ..., 1)$ in order to create a sufficiently dense test-bed set. Despite the inefficiency of this process, we follow this iterative sampling procedure for its simplicity. Operational versions of this work could be enhanced by adapting optimization methods from the full conformal methodology (Papadopoulos et al., 2011; Cherubin et al., 2021).

5. Results

Figure 3 presents the coverage results derived from both prediction intervals and confidence intervals. For both methods, coverage is defined as the fraction of ground truth samples included in the predicted interval of confidence. The conformal framework achieves validity in all the scenarios, closely approaching perfect calibration in five of them, and without assuming any underlying distribution of data. The overconfidence predicted in Figures 3(a) and 3(b) is due to the low number of calibration samples. According to Figure 2,



(a) M = 5, $\kappa = 2$

(b) M = 6, $\kappa = 3$

(c) M = 7, $\kappa = 5$

(d) M = 8, $\kappa = 4$

(e) M = 9, $\kappa = 2$

(f) M = 10, $\kappa = 6$

Figure 1: Stability estimates of the possible c subsamples, each consisting of κ elements drawn from a sample with M elements, are shown in blue. The oracle line denotes the true stability, while the black dashed line represents the average of the c subsamples.

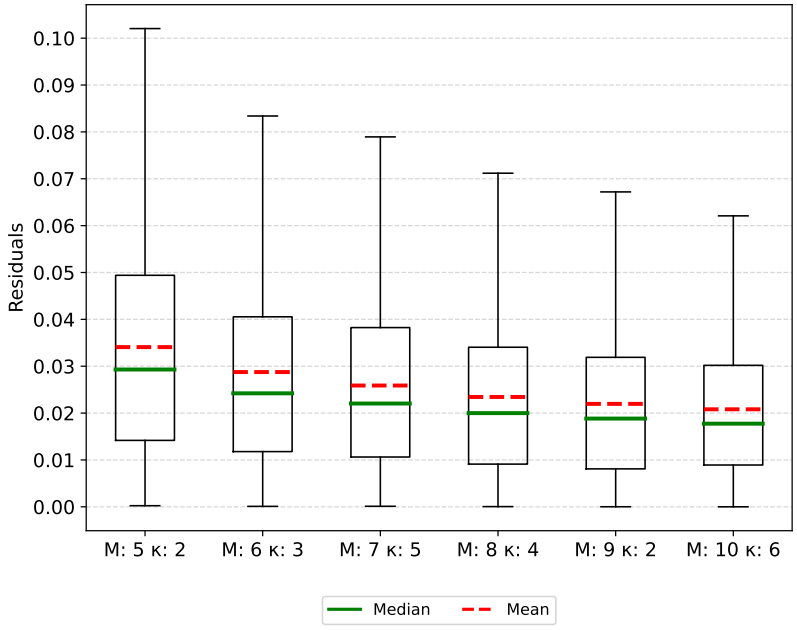


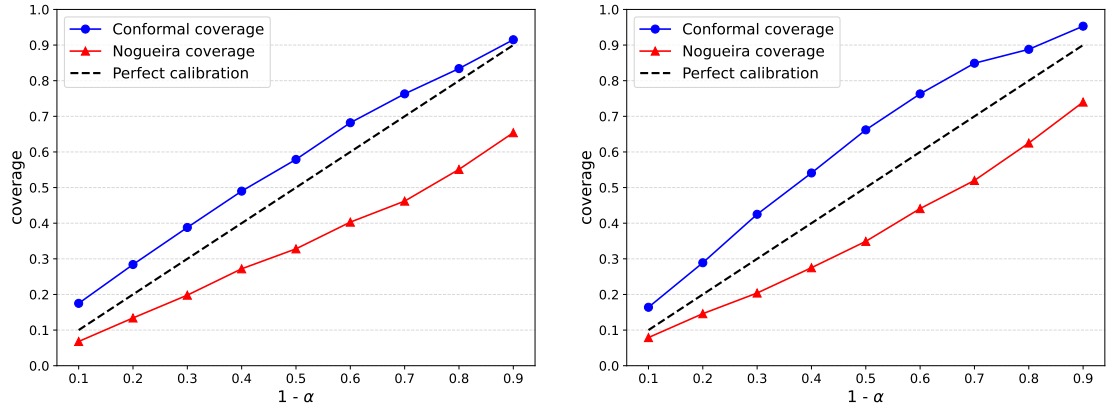
Figure 2: Boxplot with residuals of 1000 independent samples using different combinations of M and κ.

the number of points available to 'fit' the mean is too low, leading to a low efficiency for these cases. The coverage reported using the confidence intervals from Equation (3) leads to miscoverage because this only holds true when $M \to \infty$. Table 1 shows the coverage difference with respect to perfect calibration $\boldsymbol{\delta}_{Cov}$ for both prediction intervals and confidence intervals. Due to space efficiency, only the results for $1-\alpha = 0.9, 0.7, 0.5$, and 0.3 are shown.

We also evaluated the efficiency of both methods. Table 1 shows the efficiency as the difference between the endpoints of the uncertainty intervals $\boldsymbol{\Delta}_{Eff}$, averaged over the 1000 independent tests. Note that the prediction intervals are slightly larger than the confidence intervals. This effect occurs because, due to the coverage guarantees, prediction intervals must be wider. Nevertheless, in terms of informativeness, prediction intervals seem to be as efficient as confidence intervals. In addition, as M increases, the average length of prediction intervals approaches the average length of confidence intervals.

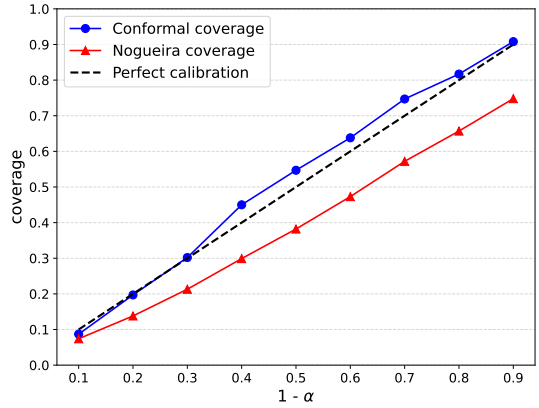
6. Conclusions

The stability of a feature selection algorithm quantifies how different sets of samples affect the selection of a relevant subset of features. This property is essential to increase confidence in the features selected by an algorithm. In this work, we present a procedure for constructing well-calibrated prediction intervals to measure the stability of a feature selection method. We focus on the situation in which only a few subsets of selected features are available after running a computationally demanding feature selection procedure. The

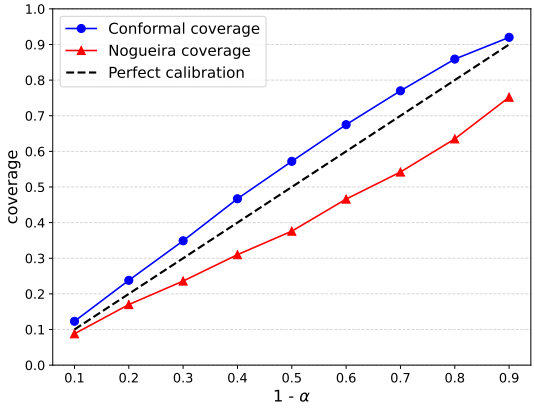


(a) M = 5

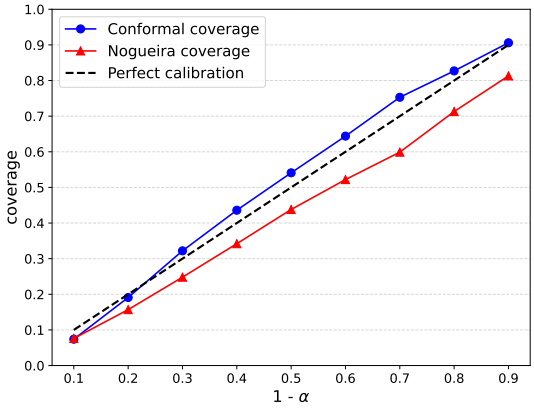
(b) M = 6



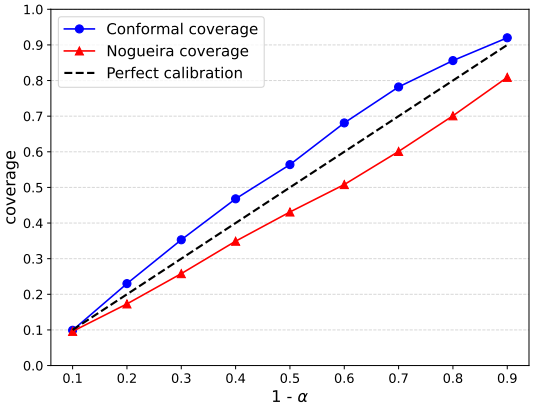
(c) M = 7



(d) M = 8



(e) M = 9



(f) M = 10

Figure 3: Coverage comparison between prediction intervals in blue and confidence intervals in red. The black dashed line represents the expected output from a perfectly calibrated model.

M	Method	1 - α = 0.9		0.7		0.5		0.3	
		Δ_{Eff}	δ_{Cov}	Δ_{Eff}	δ_{Cov}	Δ_{Eff}	δ_{Cov}	Δ_{Eff}	δ_{Cov}
5	**P.I.**	0.17(0.06)	+0.02	0.11(0.04)	+0.06	0.07(0.03)	+0.08	0.04(0.02)	+0.09
	C.I.	0.08(0.03)	-0.25	0.05(0.02)	-0.24	0.03(0.01)	-0.17	0.02(0.01)	-0.10
6	**P.I.**	0.17(0.05)	+0.05	0.11(0.03)	+0.15	0.07(0.02)	+0.16	0.04(0.02)	+0.12
	C.I.	0.08(0.03)	-0.16	0.05(0.02)	-0.18	0.03(0.01)	-0.15	0.02(0.01)	-0.10
7	**P.I.**	0.12(0.03)	+0.01	0.08(0.02)	+0.05	0.05(0.02)	+0.05	0.03(0.01)	+0.00
	C.I.	0.08(0.02)	-0.15	0.05(0.01)	-0.13	0.03(0.01)	-0.12	0.02(0.01)	-0.09
8	**P.I.**	0.12(0.03)	+0.02	0.08(0.02)	+0.07	0.05(0.01)	+0.07	0.03(0.01)	+0.05
	C.I.	0.07(0.02)	-0.15	0.05(0.01)	-0.16	0.03(0.01)	-0.12	0.02(0.00)	-0.06
9	**P.I.**	0.10(0.02)	+0.01	0.06(0.02)	+0.05	0.04(0.01)	+0.04	0.02(0.01)	+0.02
	C.I.	0.07(0.02)	-0.09	0.04(0.01)	-0.10	0.03(0.01)	-0.06	0.02(0.00)	-0.05
10	**P.I.**	0.10(0.02)	+0.02	0.06(0.01)	+0.08	0.04(0.01)	+0.06	0.02(0.01)	+0.05
	C.I.	0.07(0.02)	-0.09	0.04(0.01)	-0.10	0.03(0.01)	-0.07	0.02(0.00)	-0.04

Table 1: Efficiency $\boldsymbol{\Delta}_{Eff}$ and the deviation from perfect calibration $\boldsymbol{\delta}_{Cov}$. Efficiency is given as the average length of the prediction intervals (P.I.) and the confidence intervals (C.I.), respectively. The standard deviation is given in parentheses. Deviations from the perfect calibration correspond to the values observed in Figure 3. The + represents deviation as overcoverage, whereas − represents that the deviation is a misscoverage. A deviation of 0.00 corresponds to a deviation lower than ±0.005.

developed framework can be applied to any stability estimator, as long as the subsample stability converges to the ground truth. We specifically use Nogueira's estimator to measure stability (3). This estimator fulfils the desirable properties that a stability estimator should have and also allows the construction of approximate confidence intervals. Our empirical work showed that the conformal prediction intervals achieves both validity and efficiency in a scenario with a low number of samples, whereas the confidence intervals provided by the Nogueira's framework are not valid. This is due to the finite sample validity property in conformal prediction.

The main limitation of our work is related to the simplicity of the estimator and the defined non-conformity measure. As Figure 3 shows, despite the validity achieved by the prediction intervals, their efficiency can be improved when the number of samples available is low. Future research should investigate and implement more sophisticated methods, such as fitting a kernel density estimator (Jing Lei, 2014). Further research should also explore this procedure in real-world applications, other stability measures and estimators, and the inductive conformal prediction framework. Finally, wider prediction intervals may

be a multiplicity indicator; even so stability theory must be adapted and new estimators developed to overcome this challenge.

Acknowledgments

This work was partially supported by the Gobierno de Navarra through the ANDIA 2021 program (grant no. 0011-3947-2021-000023) and the ERA PerMed JTC2022 PORTRAIT project (grant no. 0011-2750-2022-000000). We would like to thank Nick Guthrie for his valuable input with regard to language editing.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundantions and Trends in Machine Learning*, 16(4):494–591, 2023.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Morgan Kaufmann, San Francisco, CA, USA, 2014.

Concha Bielza and Pedro Larrañaga. *Data-Driven Computational Neuroscience: Machine Learning and Statistical Models*. Cambridge University Press, Cambridge, U.K., 2020.

Leo Breiman. Random forests. *Machine Learning*, (45):5–32, 2001.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012.

Giovanni Cherubin, Konstantinos Chatzikokolakis, and Martin Jaggi. Exact optimization of conformal predictors via incremental and decremental learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1836–1845. PMLR, 2021.

Chad A. Davis, Fabian Gerick, Volker Hintermair, Caroline C. Friedel, Katrin Fundel, Robert Küffner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Alex Gammerman, Vladimir Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

Isabelle Guyon, Masoud Nikravesh, Steve Gunn, Lotfi A. Zadeh, and Janusz Kacprzyk, editors. *Feature Extraction: Foundations and Applications*. Springer, Berlin, Heidelberg, 2006. ISBN 978-3-540-35487-1 978-3-540-35488-8.

Kilem Li Gwet. Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika*, 73:407–430, 2008.

Larry Wasserman Jing Lei. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Giuseppe Jurman, Stefano Merler, Annalisa Barla, Silvano Paoli, Antonio Galea, and Cesare Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, 2008.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM)*, page 8, Houston, TX, USA, 2005.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge Information Systems*, 12: 95–116, 2007.

Ludmila I. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, page 390–395, USA, 2007. ACTA Press.

Sarah Nogueira. *Quantifying the Stability of Feature Selection*. PhD thesis, University of Manchester, 2018.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.

H. Papadopoulos, V. Vovk, and A. Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, page 345–356, Berlin, Heidelberg, 2002. Springer-Verlag.

Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Alexander Statnikov and Constantin F. Aliferis. Analysis and computational dissection of molecular signature multiplicity. *PLOS Computational Biology*, 6(5):1–9, 2010.

Alexander Statnikov, Nikita I. Lytkin, Jan Lemeire, and Constantin F. Aliferis. Algorithms for Discovery of Multiple Markov Boundaries. *Journal of Machine Learning Research*, 14:499–566, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Paolo Toccaceli. Introduction to conformal predictors. *Pattern Recognition*, 124:108507, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer New York, NY, 1 edition, 2005.