Conformal Stability Measure for Feature Selection Algorithms

2024 Conformal and Probabilistic Prediction with Applications

9th to 11th September - Milano, Italy

Marcos López-De-Castro (PhD student),

Alberto García-Galindo, Rubén Armañanzas

Outline

1. Motivation
2. The framework
3. The approach based on CP
4. Results
5. Conclusions, limitations and further work

Outline

Motivation

- Why are we interested in Feature Selection?

Motivation

- Why are we interested in Feature Selection? $\rightarrow$ Knowledge discovery.

Motivation

- Why are we interested in Feature Selection? $\rightarrow$ Knowledge discovery.

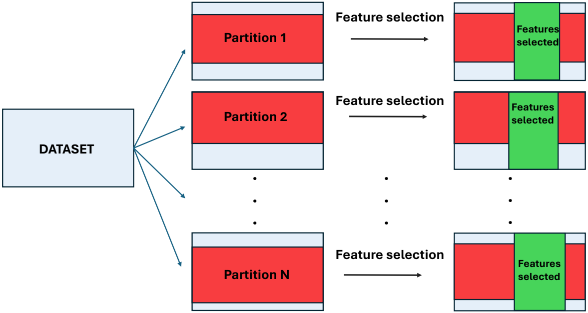


Figure 1: Stability of Features.

Motivation

- Why are we interested in Feature Selection? → Knowledge discovery.

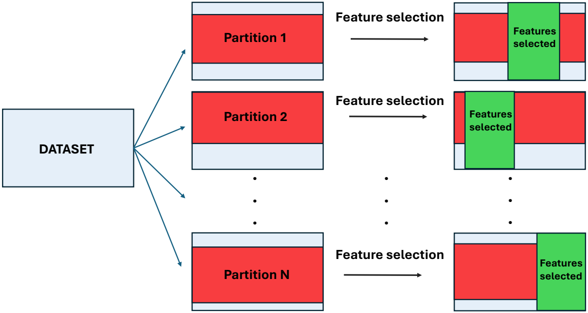


Figure 2: Non-stability of Features.

Outline

Previous works on stability

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

The framework

- Stability $\iff$ RV.
- Estimator has 5 desirable properties:

The framework

- Stability $\iff$ RV.
- Estimator has 5 desirable properties:
  - Fully defined
  - Strict monotonicity
  - Known bounds
  - Maximum stability if and only if the selection is deterministic
  - Correction for chance

The framework

- Stability $\iff$ RV.
- Estimator has 5 desirable properties:
  - Fully defined
  - Strict monotonicity
  - Known bounds
  - Maximum stability if and only if the selection is deterministic
  - Correction for chance
- Framework computes approximate Confidence Intervals (CI) for the estimates:

The framework

- Stability $\iff$ RV.
- Estimator has 5 desirable properties:
    - Fully defined
    - Strict monotonicity
    - Known bounds
    - Maximum stability if and only if the selection is deterministic
    - Correction for chance
- Framework computes approximate Confidence Intervals (CI) for the estimates:
    - **No valid guarantees**
    - **Only empirical approaches asymptotically**

The framework

- Stability $\iff$ RV.
- Estimator has 5 desirable properties:
  - Fully defined
  - Strict monotonicity
  - Known bounds
  - Maximum stability if and only if the selection is deterministic
  - Correction for chance
- Framework computes approximate Confidence Intervals (CI) for the estimates:
  - **No valid guarantees**
  - **Only empirical approaches asymptotically**
- **Our Contribution:**
  - Use Conformal Prediction (CP) to provide valid and non-asymptotic prediction intervals of stability.

The framework

- $\mathcal{D} = \{(X, Y)\}$, where $X \in \mathbb{R}^d$.
- Let $\pi(\cdot)$ be a feature selection method.

The framework

- $\mathcal{D} = \{(X, Y)\}$, where $X \in \mathbb{R}^d$.
- Let $\pi(\cdot)$ be a feature selection method.
- $\pi(D) = z$ where z is a binary string of length d,

$$z = (0, 1, 1, 0, 0, 1).$$

The framework

- $\mathcal{D} = \{(X, Y)\}$, where $X \in \mathbb{R}^d$.
- Let $\pi(\cdot)$ be a feature selection method.
- $\pi(D) = z$ where z is a binary string of length d,

$$z = (0, 1, 1, 0, 0, 1).$$

- If we take M bootstrap samples from $\mathcal{D} \to$ matrix $\mathcal{Z}$:

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

The framework

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

- **1st key assumption:** We assume independence between the rows of matrix $\mathcal{Z}$.

The framework

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

- **2nd key assumption:** Under the 1st assumption, the columns of matrix $\mathcal{Z}$ are random variables following a *Bernouilli distribution* with mean parameters b_j.

The framework: Nogueira's estimator

Definition (Stability estimator)

A stability estimator for feature selection algorithms is as follows:

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

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* $\mathcal{Z}$.

The framework: Nogueira's estimator

Definition (Stability estimator)

A stability estimator for feature selection algorithms is as follows:

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

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* $\mathcal{Z}$.

Definition ($\hat{\Phi}_N$ confidence interval)

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

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

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.*

Outline

The approach based on CP

- Marginal coverage:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})) \geq 1 - \alpha, \tag{3}$$

- We want

$$\mathbb{P}(\Phi \in \mathcal{C}_\alpha(\mathcal{Z})) \geq 1 - \alpha. \tag{4}$$

The approach based on CP: Methodology

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

$$\begin{pmatrix} 0 & 1 & 0 & \ddots & 1 \\ 1 & 1 & 0 & \cdots & 0 \end{pmatrix}_{\kappa \times d}, \qquad \cdots \qquad \begin{pmatrix} 1 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{\kappa \times d}.$$

Subsampling of the matrix $\mathcal{Z}$ by rows.
A set $\mathcal{R} = \{\mathcal{Z}_1, ..., \mathcal{Z}_c\}$ is generated.
$\mathcal{Z}_i$ is a $\kappa \times d$ binary matrix with $\kappa < M$.

The approach based on CP: Methodology

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \end{pmatrix}, \qquad \cdots \qquad \begin{pmatrix} 1 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}.$$

- **Independence between rows of $\mathcal{Z}$.**
- **Columns of $\mathcal{Z}$ follows** $\mathcal{B}(b_j)$.

The approach based on CP: Methodology

$$\mathcal{Z} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}_{M \times d}$$

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \end{pmatrix}, \qquad \cdots \qquad \begin{pmatrix} 1 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}.$$

- **Independence between rows of $\mathcal{Z}$.**
- **Columns of $\mathcal{Z}$ follows** $\mathcal{B}(b_j)$.

$\Rightarrow$ Compute stabilities of elements in $\mathcal{R}$. $\Rightarrow$ **Indistinguishable.**

The Approach Based on CP: Methodology

- $\{\hat{\Phi}_N(\mathcal{Z}_1), \ldots, \hat{\Phi}_N(\mathcal{Z}_i), \ldots, \hat{\Phi}_N(\mathcal{Z}_c)\} \leftarrow$ **Bag of samples** $\mathcal{R}$

The Approach Based on CP: Methodology

- $\{\hat{\Phi}_N(\mathcal{Z}_1), \ldots, \hat{\Phi}_N(\mathcal{Z}_i), \ldots, \hat{\Phi}_N(\mathcal{Z}_c)\} \leftarrow$ **Bag of samples** $\mathcal{R}$

Transductive CP Algorithm:

- **Initialize:**
  - Define a point estimate $\hat{\theta}_z$ based on the bag.
  - Define $f()$: the distance between the point estimate and a sample.
  - Propose a set of trial values $\hat{\Phi}_N(z) \in \mathcal{Z}_{trial} = \{-\frac{1}{\kappa-1}, ..., 1\}$.

The Approach Based on CP: Methodology

- $\{\hat{\Phi}_N(\mathcal{Z}_1), \ldots, \hat{\Phi}_N(\mathcal{Z}_i), \ldots, \hat{\Phi}_N(\mathcal{Z}_c)\} \leftarrow$ **Bag of samples** $\mathcal{R}$

Transductive CP Algorithm:

- **Initialize:**
  - Define a point estimate $\hat{\theta}_z$ based on the bag.
  - Define $f()$: the distance between the point estimate and a sample.
  - Propose a set of trial values $\hat{\Phi}_N(z) \in \mathcal{Z}_{trial} = \{-\frac{1}{\kappa-1}, ..., 1\}$.
- **Compute Non-conformity Measures:**

$$\varphi_{z,i} = f(\hat{\theta}_z, \hat{\Phi}_N(\mathcal{Z}_i)) \quad \forall i \in \{1, \ldots, c\},$$
$$\varphi_{z,c+1} = f(\hat{\theta}_z, \hat{\Phi}_N(z))$$

The Approach Based on CP: Methodology

- $\{\hat{\Phi}_N(\mathcal{Z}_1), \ldots, \hat{\Phi}_N(\mathcal{Z}_i), \ldots, \hat{\Phi}_N(\mathcal{Z}_c)\} \leftarrow$ **Bag of samples** $\mathcal{R}$

Transductive CP Algorithm:

- **Initialize:**
  - Define a point estimate $\hat{\theta}_z$ based on the bag.
  - Define $f()$: the distance between the point estimate and a sample.
  - Propose a set of trial values $\hat{\Phi}_N(z) \in \mathcal{Z}_{trial} = \{-\frac{1}{\kappa-1}, ..., 1\}$.
- **Compute Non-conformity Measures:**

$$\varphi_{z,i} = f(\hat{\theta}_z, \hat{\Phi}_N(\mathcal{Z}_i)) \quad \forall i \in \{1, \ldots, c\},$$
$$\varphi_{z,c+1} = f(\hat{\theta}_z, \hat{\Phi}_N(z))$$

- **Check Conformity:**

$$\mathcal{C}_\alpha \leftarrow \{\hat{\Phi}_N(z)_j \in \mathcal{Z}_{trial} : p^j > \alpha\}$$

Outline

Design of experiments

- **Tests:** Artificial datasets codified as $\mathcal{Z}$.
  - $M \times 100$ binary matrix $\mathcal{Z}$ with $M = m, \ \forall m \in \{5, ..., 10\}$.
  - Columns are drawn from $\mathcal{B}(b_j)$, with known b_j (so the true stability is known).
  - We performed 1000 independent simulations for each m.
  - 500 test values equally-spaced.

Design of experiments

- **Tests:** Artificial datasets codified as $\mathcal{Z}$.
  - $M \times 100$ binary matrix $\mathcal{Z}$ with $M = m, \ \forall m \in \{5, ..., 10\}$.
  - Columns are drawn from $\mathcal{B}(b_j)$, with known b_j (so the true stability is known).
  - We performed 1000 independent simulations for each m.
  - 500 test values equally-spaced.

- **Non-conformity score:**

$$\varphi_{z,i} = \left|\frac{\hat{\Phi}(\mathcal{Z}_i) - \mu_z}{\sigma_z}\right|, \tag{5}$$

where μ_z, σ_z are the mean and the standard deviation of $\mathcal{R} \cup \{z\} - \{\hat{\Phi}(\mathcal{Z}_i)\}$ and z is a trial value.

Some results

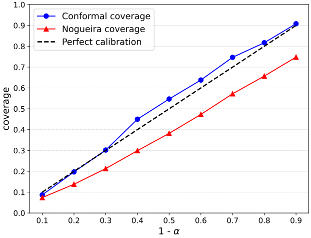


Figure 3: M = 7

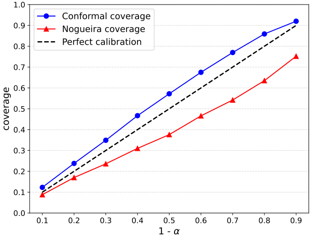


Figure 4: M = 8

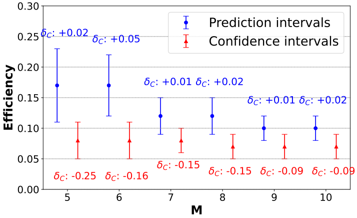


Figure 5: $1 - \alpha = 0.9$

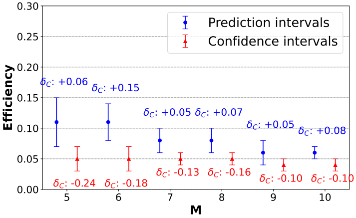


Figure 6: $1 - \alpha = 0.7$

Outline

1. Motivation
2. The framework
3. The approach based on CP
4. Results
5. Conclusions, limitations and further work

Conclusions, limitations and further work

- **Conclusions:**
  - Well-calibrated prediction intervals to estimate the stability of any feature selection method.
  - Prediction intervals achieves validity and efficiency converges to C.I.

Conclusions, limitations and further work

- **Conclusions:**
  - Well-calibrated prediction intervals to estimate the stability of any feature selection method.
  - Prediction intervals achieves validity and efficiency converges to C.I.
- **Limitations:**
  - Improve efficiency when the number of samples available is low.
  - May be computationally demanding (iterative sampling procedure).

Conclusions, limitations and further work

- **Conclusions:**
  - Well-calibrated prediction intervals to estimate the stability of any feature selection method.
  - Prediction intervals achieves validity and efficiency converges to C.I.
- **Limitations:**
  - Improve efficiency when the number of samples available is low.
  - May be computationally demanding (iterative sampling procedure).
- **Future work:**
  - Define better point estimators.
  - New non-conformity functions.
  - 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).
  - Extension to split CP?

Thanks for your attention!!

Acknowledgements

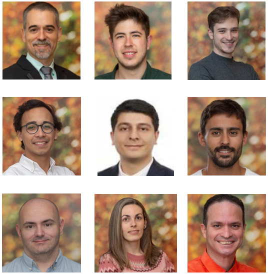
Gobierno de Navarra | Nafarroako Gobernua | AGENDA 2030

European Commission

Horizon 2020
European Union funding
for Research & Innovation

ERA PerMed

Universidad de Navarra | DATAI
INSTITUTO DE CIENCIA DE LOS
DATOS E INTELIGENCIA ARTIFICIAL



Contact: *mlopezdecas@unav.es*

COPA

Universidad de Navarra | DATAI INSTITUTO DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

Conformal Stability Measure for Feature Selection Algorithms

2024 Conformal and Probabilistic Prediction with Applications

9th to 11th September - Milano, Italy

Marcos López-De-Castro (PhD student),

Alberto García-Galindo, Rubén Armañanzas