Introduction to Conformal Prediction with Application on Imbalanced Diabetes Classification

Mabrouka Salmi

Master en Inteligencia Computacional e IOT

Escuela Politécnica Superior de Córdoba

Universidad de Córdoba - UCO, Córdoba, España

April 18, 2024

Motivation

- Traditional ML produces bare predictions

Points prediction, ex: diabetes prediction based on a set of features (Pima Indian Diabetes dataset).

```
print(model.predict(x_test)):  ['diabetes']
print(model.predict_proba(x_test)):  [[0.8 0.2]]
```

Motivation

- Traditional ML produces bare predictions
- How good is the prediction?

How close is that to the true value?

Can we trust the model?

Motivation

- Traditional ML produces bare predictions
- How good is the prediction?
- Anticipating past performance on unseen/future data

The accuracy on test data 77% –> we assume similar accuracy on production data

How confident are the model's predictions on new data?

Motivation

- Traditional ML produces bare predictions
- How good is the prediction?
- Anticipating past performance on unseen/future data
- Estimating the prediction uncertainty

To quantify the uncertainty, using sets/intervals that "likely contain the true prediction"

(Frequentist/Bayesian Approach: uncalibrated, data distribution/prior assumptions)

Conformal Prediction

- Per-instance confidence estimates: Interval/set predictions

Conformal Prediction

- Per-instance confidence estimates: Interval/set predictions
- Not reliant on prior probabilities / non-parametric

Conformal Prediction

- Per-instance confidence estimates: Interval/set predictions
- Not reliant on prior probabilities / non-parametric
- Model Agnostic

Conformal Prediction

- Per-instance confidence estimates: Interval/set predictions
- Not reliant on prior probabilities / non-parametric
- Model Agnostic
- The approach is thoroughly validated and straightforward to put into practice.

Intuition behind it

Exchangeability

Assume that calibration and test data are exchangeable; the joint probability distribution is not dependent on the order of its arguments.

Non-conformity measures (NCM)

It assesses how much out-of-place the example appears (not conform to a collection of samples)
The randomness is judged based on the proportion of the examples having larger values of NCM than the example in hand;
Low value: rare examples look more out-of-place
High value: Most examples look more out-of-place

Intuition behind it

The procedure of prediction (classification)

- A hypothetical example is created of the test example and the hypothetical label: a hypothetical completion

Intuition behind it

The procedure of prediction (classification)

- A hypothetical example is created of the test example and the hypothetical label: a hypothetical completion
- Assessment of the randomness of the hypothetical completion

Intuition behind it

The procedure of prediction (classification)

- A hypothetical example is created of the test example and the hypothetical label: a hypothetical completion
- Assessment of the randomness of the hypothetical completion
- Include labels where the degree of randomness relative to the training set is higher than a chosen significance level α

α is user defined

Validity & efficiency

Validity property

- How similar is the hypothetical completion to the existing data in the dataset

Validity & efficiency

Validity property

- How similar is the hypothetical completion to the existing data in the dataset
- The p-value: determine the fraction of data points in the training dataset, including the hypothetical completion itself, whose NCM is greater than or equal to the hypothetical completion's NCM

Validity & efficiency

Validity property

- How similar is the hypothetical completion to the existing data in the dataset
- The p-value: determine the fraction of data points in the training dataset, including the hypothetical completion itself, whose NCM is greater than or equal to the hypothetical completion's NCM
- The prediction set includes possible label values for which the p-values exceed a chosen significance level.

Validity & efficiency

Validity property

- How similar is the hypothetical completion to the existing data in the dataset
- The p-value: determine the fraction of data points in the training dataset, including the hypothetical completion itself, whose NCM is greater than or equal to the hypothetical completion's NCM
- The prediction set includes possible label values for which the p-values exceed a chosen significance level.
- An error occurs when the prediction set fails to contain the actual label

Validity & efficiency

Validity property

- How similar is the hypothetical completion to the existing data in the dataset
- The p-value: determine the fraction of data points in the training dataset, including the hypothetical completion itself, whose NCM is greater than or equal to the hypothetical completion's NCM
- The prediction set includes possible label values for which the p-values exceed a chosen significance level.
- An error occurs when the prediction set fails to contain the actual label
- CPs exhibit conservative asymptotic validity: error rates converge a.s to values less than/equal to the significance level.

Validity & efficiency

Efficiency

- The prediction sets can contain any subset of the label space Y, including the entire label space Y and the empty set

Validity & efficiency

Efficiency

- The prediction sets can contain any subset of the label space Y, including the entire label space Y and the empty set
- Makes the prediction sets as small as possible

Validity & efficiency

Efficiency

- The prediction sets can contain any subset of the label space Y, including the entire label space Y and the empty set
- Makes the prediction sets as small as possible
- Depends on the choice of NCM.

Transductive CP

Select a non-conformity Measure

- The inverse probability or the hinge loss: $1 - P(y|x)$, $P(y|x)$ is the classification score for the actual class.

Transductive CP

Select a non-conformity Measure

- The inverse probability or the hinge loss: $1 - P(y|x)$, $P(y|x)$ is the classification score for the actual class.
- How much far is the classification score produced by our model to the probability score produced by the ideal classifier.

Transductive CP

Train the underlying classifier model

- Train the model $2 \times m$, m : the number of test instances.
  Computacionally expensive for huge datasets (refer to ICP)
  Better prediction intervals for medium and small datasets

Compute nonconformity scores for training set

- For both models for the training set augmented with the given test point

Transductive CP

Train the underlying classifier model

- Train the model $2 \times m$, m : the number of test instances.
  Computacionally expensive for huge datasets (refer to ICP)
  Better prediction intervals for medium and small datasets

Compute nonconformity scores for training set

- For both models for the training set augmented with the given test point
- One test point at a time

Transductive CP

Compute the p-values

- Compute the p-value for each test point twice (2 classes)

Construct the predictions set

- Include or exclude potential labels H_0 : *the label can be included*

Transductive CP

Compute the p-values

- Compute the p-value for each test point twice (2 classes)
- How well it fits the training set

Construct the predictions set

- Include or exclude potential labels H_0 : *the label can be included*
- If the p-value for the corresponding hypothetical label is larger than the significant level, we include such a label

Transductive CP

Compute the p-values

- Compute the p-value for each test point twice (2 classes)
- How well it fits the training set

Construct the predictions set

- Include or exclude potential labels H_0 : *the label can be included*
- If the p-value for the corresponding hypothetical label is larger than the significant level, we include such a label
- Empty set of predictions refers to an anomaly or a prediction error.

Inductive CP

- Split the data into training, calibration, and test

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)
- Choose a nonconformity measure

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)
- Choose a nonconformity measure
- Compute the nonconformity scores

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)
- Choose a nonconformity measure
- Compute the nonconformity scores
- Compute the P-values for each instance of the calibration dataset

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)
- Choose a nonconformity measure
- Compute the nonconformity scores
- Compute the P-values for each instance of the calibration dataset
- Fix a significance level in [0,1]

Inductive CP

- Split the data into training, calibration, and test
- Train the underlying classifier model on training data (only once)
- Choose a nonconformity measure
- Compute the nonconformity scores
- Compute the P-values for each instance of the calibration dataset
- Fix a significance level in [0,1]
- Include(or exclude) the hypothetical label: if the p-value is larger than the significance level, include, Else, exclude.

Label Conditional CP

Label-conditional conformal prediction incorporates label information into the nonconformity scoring process.

- Computing the nonconformity scores is conditioned on the predicted label
- More accurate prediction in class imbalance contexts.

Methodology

- Research question: How does conformal prediction perform in classifying imbalanced medical data compared to traditional machine learning methods?
- Pima Indian Diabetes Dataset
- Three CP algorithms: Transductive, Inductive, Label-Conditional Classifier
- Base classifier Random Forest
- Evaluation metrics: error rate (validity), average number of classes per prediction (efficiency), and accuracy

Preliminary results

Table 1: Conformal prediction models results on Pima diabetes dataset

CP Model\Metric	**Error Rate**	**Avg N of classes**
Inductive CP	0.1034	1.30
Transductive CP	0.1168	1.3311
Label-Conditional CP	0.1168	1.3831

Preliminary results

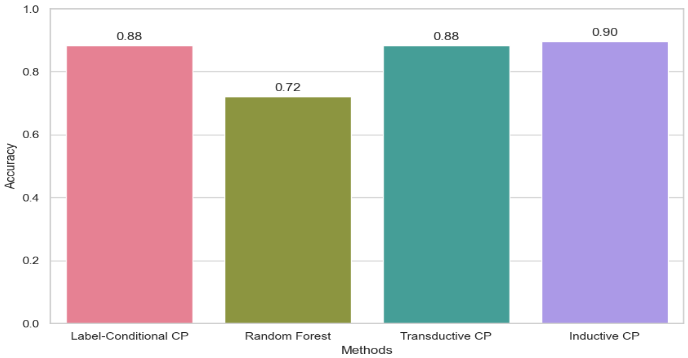

Figure 1: Comparison of model accuracies for diabetes classification

Discussion

- CP model exceeds the non-CP model in terms of accuracy.
- CP models achieve optimal accuracy, in particular, Inductive CP with the lowest error rate
- The obtained classification accuracy is better than that of achieved in a recent survey of ML algorithms for Pima Indian Diabetes (accuracy up to 80%)
- Further assessment of the predictions could inform more on the efficiency of the CP models.

1. Introduction
2. Conformal Prediction
3. Methodology & Experimental work
4. **Conclusions & Future Work**

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
    - An initial evaluation of three CP models against non-CP.

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases
  - Promising results on moderately imbalanced and small medical data

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
    - An initial evaluation of three CP models against non-CP.
    - CP models improved the prediction accuracy of diabetes cases
    - Promising results on moderately imbalanced and small medical data

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases
  - Promising results on moderately imbalanced and small medical data
- Future work

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
    - An initial evaluation of three CP models against non-CP.
    - CP models improved the prediction accuracy of diabetes cases
    - Promising results on moderately imbalanced and small medical data
- Future work
    - Comprehensive analysis of the CP algorithms on imbalanced medical data with different features.

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases
  - Promising results on moderately imbalanced and small medical data
- Future work
  - Comprehensive analysis of the CP algorithms on imbalanced medical data with different features.
- Review on the original work

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases
  - Promising results on moderately imbalanced and small medical data
- Future work
  - Comprehensive analysis of the CP algorithms on imbalanced medical data with different features.
- Review on the original work
  - Preliminary review on CP

- This study presents a preliminary exploration of Conformal Prediction (CP) models in the classification of imbalanced medical data
  - An initial evaluation of three CP models against non-CP.
  - CP models improved the prediction accuracy of diabetes cases
  - Promising results on moderately imbalanced and small medical data
- Future work
  - Comprehensive analysis of the CP algorithms on imbalanced medical data with different features.
- Review on the original work
  - Preliminary review on CP
  - Lack of an experimental/ thorough review; update with the advancements; consensus on best practices for CP

Thank you!