Confidence Classifiers with Guaranteed Accuracy or Precision

COPA 2023

Ulf Johansson[1], Cecilia Sönströd[1], Tuwe Löfström[1], Henrik Boström[2]

September 11, 2023

[1]Jönköping University, Sweden.
Email: {ulf.johansson, cecilia.sonstrod, tuwe.lofstrom}@ju.se

[2]KTH Royal Institute of Technology, Sweden. Email: bostromh@kth.se

Summary - TLDR

- We suggest and evaluate a new algorithm for **classification with reject option**.
- The overall idea is that the algorithm **estimates accuracy or precision for different rejection levels**.
- The basis of the suggested algorithm is **conformal classification**, so it comes with **validity guarantees**.
- The experimentation, using 10 publicly available two-class data sets, confirms that the precision and accuracy estimates are **excellent**.
- In an outright comparison with **probabilistic predictors**, including models calibrated with Platt scaling, the suggested algorithm **clearly outperforms** the alternatives.

Outline

1. Classification with reject option
2. Conformal classification
3. Suggested approach
4. Method
5. Results
6. Conclusions

Classification with reject option

Classification with reject option

- A standard classifier is forced to predict the label of every test instance, even when confidence in the predictions is very low.
- When the model is used for **decision support**, either by a human decision-maker, or as part of an automated system, this poses problems.
- In many scenarios, it would actually be better to **avoid** making these predictions.
- A classifier with that alternative is referred to as a **classifier with reject option**.
- A real-world example is predicting **product returns** in e-tailing.
- Here, however, a **false positive** (acting on an order that would not have been returned) is much worse than missing an order that will be returned.
- So, for that specific problem, **precision** is more important than overall accuracy.

Classification with reject option

- A typical scenario for using classification with reject option is to refer instances rejected by the model to a **human expert**, for manual decisions, possibly aided by the model and/or accompanying explanations.
- In these cases, the **trade-off** between **rejection rate** and **accuracy or precision** is a key issue, since there can be costs associated both with errors and with human processing.
- Having access to **well-calibrated** accuracy or precision estimates for different rejection rates would be extremely valuable in this situation.
- Showing how to obtain exactly that, i.e., **perfectly calibrated accuracy or precision estimates** for **different rejection rates**, by using **conformal classification**, is the overall purpose of this paper.

Conformal classification

Conformal classification

- Conformal prediction utilizes **nonconformity functions** $A : X \times Y \rightarrow \mathbb{R}$ for measuring the relative strangeness of an instance $(\boldsymbol{x}, y)$ compared to a set of instances with known target values.
- In conformal classification, a test instance is tentatively labeled $(\boldsymbol{x}_{k+1}, \tilde{y})$, and then a p-value statistic is calculated from the nonconformity scores to attempt to reject the hypothesis that $\tilde{y}$ is the true label y_{k+1} at a chosen significance level ϵ.
- This procedure is repeated for all possible labels, resulting in a **set of labels** $\tilde{y} \subseteq Y$ that were not rejected.
- This set, **per construction and under exchangeability**, contains the true target y_{k+1} with a probability of $1 - \epsilon$.

In classification, the nonconformity function is most often based on the prediction error of an underlying machine learning model. The option used in this study is the **hinge loss**.

$$\Delta\left[h\left(\boldsymbol{x}_i\right), \tilde{y}\right] = 1 - \hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}_i\right), \tag{1}$$

where $\hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}\right)$ is the probability estimate provided by the machine learning model h that the instance $\boldsymbol{x}_i$ has label $\tilde{y}$.

The problem with conformal classification

- While the theory behind conformal classification is **solid**, and the guarantees **strong**, it is **awkward** to utilize conformal classifiers and their set predictions for decision making.
- Specifically, a decision-maker would most likely be tempted to focus on **singleton** predictions, i.e., label sets containing only one label.
- In a classification with reject option scenario, one strategy could be to reject everything but singleton predictions.
- However, when looking at the singleton predictions, it is not straightforward to estimate the probability that these are in fact correct.
- In particular, the probability of a singleton prediction being an error is almost guaranteed to be (significantly) higher than ϵ.

The problem with conformal classification

- The guarantees only hold **a priori**; once a predicted label set has been observed, the likelihood of an error is highly dependent on the **size** of the label set.
- If the label set contains all labels, it **cannot** be an error; and if it is empty, it **must** be an error.
- Looking at two-class problems, and the quite frequent situation where there are no (or very few) empty predictions, **all** (or almost all) errors most come from the singleton predictions.
- With this in mind, researchers and practitioners have moved away from conformal classification, instead using probabilistic predictors, including **Venn predictors**.

Suggested approach

Confidence - credibility measures

We can, in addition to create set predictions, use the p-values to, for each test instance $\boldsymbol{x}_j$, calculate the following two values:

- The **confidence**, calculated as one minus the second largest p-value.
- The **credibility**, which is equal to the largest p-value.

The connections between confidence-credibility measures, and the set predictions are:

- The confidence is the **highest** significance level where we get a **singleton prediction**.
- The credibility is the **lowest** significance level where **all labels are rejected**.

In this paper, we will utilize **the confidence measure, and a set of test instances**, to produce classifiers with reject option having **statistical guarantees**.

- To appreciate the approach, it is vital to understand exactly what a confidence score λ_j for the test instance x_j represents.
- The correct interpretation is that if we look at **all** m predictions with a confidence of **at least** λ_j these will contain (on average) $n(1 - \lambda_j)$ errors, where n is the **total** number of predictions made, i.e., the size of the test set. (See the section *Probabilities vs. p-values*, pp. 162-163, ALRW, first edition.)

In this synthethic example, we should expect approximately three errors in total, two errors among the predictions for idx 2-9 and one error among the predictions for idx 5-9.

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
λ (confidence)	**0.70**	0.75	**0.80**	0.83	0.87	**0.90**	0.93	0.95	0.97	0.99

In a classifier with reject scenario, if we reject all instances with a confidence score lower than λ_j, the expected error rate of the predicted instances is:

$$\frac{n \cdot (1 - \lambda_j)}{m} \tag{2}$$

- If the conformal classifier is a standard ICP, the guarantees will be for overall error rate, i.e., accuracy.
- If a Mondrian setup is used, the guarantees will apply to each category individually.
- In this paper, we suggest using a Mondrian taxonomy where the categories are determined from the label predicted by the underlying model.
- With this setup, we get guarantees for the predictions of the positive class (label 1), i.e., precision.

Method

4

Approaches evaluated

Underlying models

- Decision trees
- Random forests - 300 trees.

Setups

- Conformal (C): A conformal classifier is generated and calibrated on a calibration set, before predicting the test set and producing confidence scores.
- Platt (P): The scores from the underlying model are calibrated on the same calibration set as for (C) using Platt scaling.
- Uncalibrated (U): The scores from the underlying model are used without external calibration.

Experimentation setup and data sets

- All experimentation was performed using **scikit-learn**.
- **Repeated hold-out**, with 100 repetitions using 75/25 split.
- Proper training set 2/3 of the training instances, calibration set 1/3.
- **Rejection rates** $\in \{0.1, 0.2, \ldots, 0.9\}$

Data set	#inst	#attrib	prop. pos.	Source
creditA	690	16	0.44	UCI
diabetes	768	9	0.35	UCI
german	1000	21	0.70	UCI
kc1	2109	22	**0.26**	Promise
kc2	522	22	**0.27**	Promise
kr-vs-kp	3196	36	0.52	UCI
pc4	1458	38	**0.13**	Promise
transfusion	748	5	**0.26**	UCI
tic-tac-toe	958	10	0.65	UCI
wbc	699	10	0.35	UCI

Results

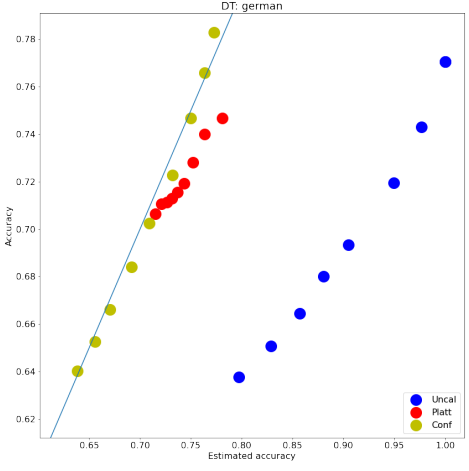
5

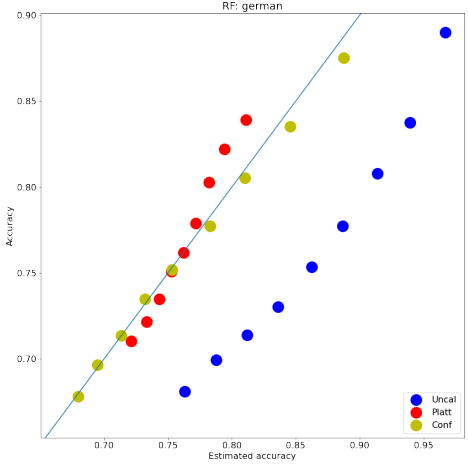
Detailed results can be found in the paper.

Here, I will only show some examples.

Example: German data set (accuracy)

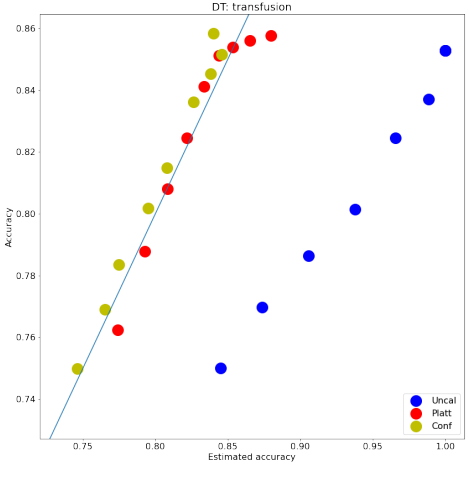
Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated accuracy and the y-axis as the actual accuracy obtained.

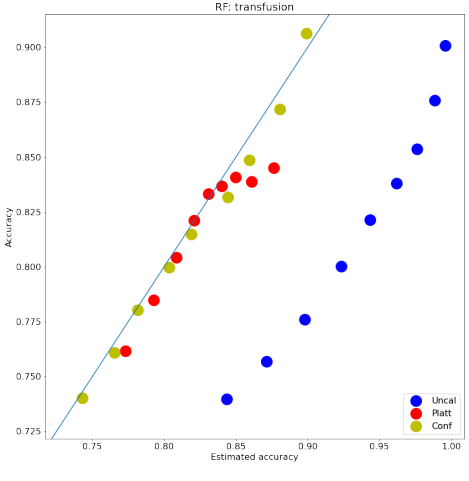




Example: Transfusion data set (accuracy)

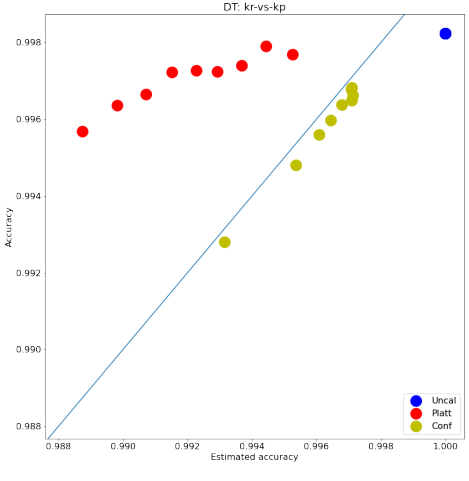
Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated accuracy and the y-axis as the actual accuracy obtained.

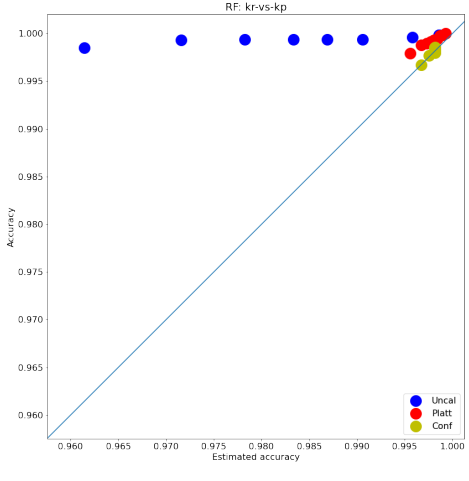




Example: Kr-vs-kp data set (accuracy)

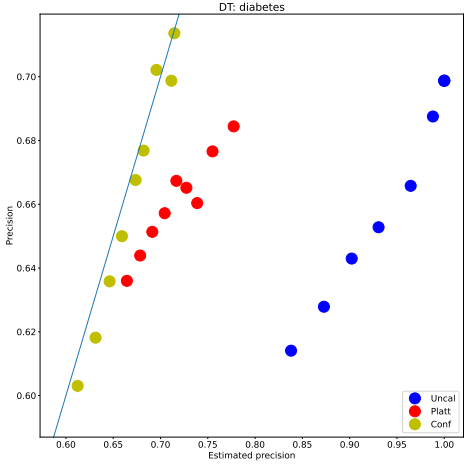
Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated accuracy and the y-axis as the actual accuracy obtained.

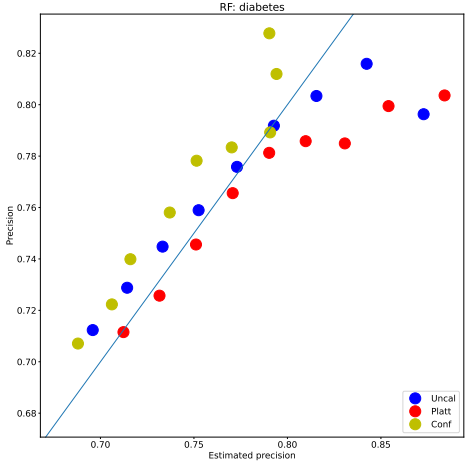




Example: Diabetes data set (precision)

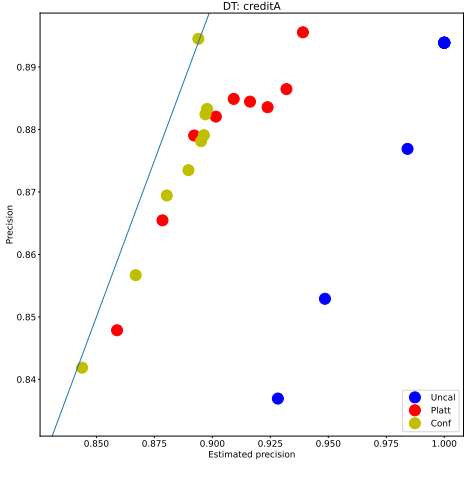
Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated precision and the y-axis as the actual precision obtained.

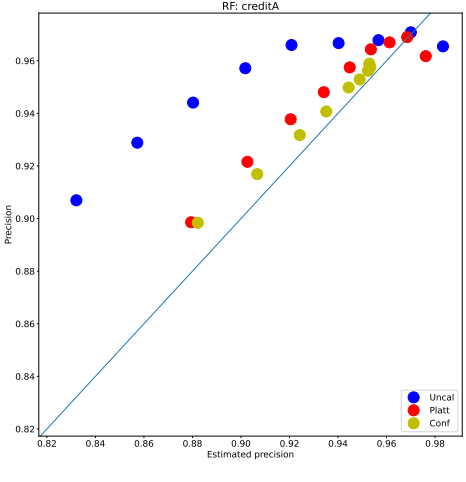




Example: Credit-A data set (precision)

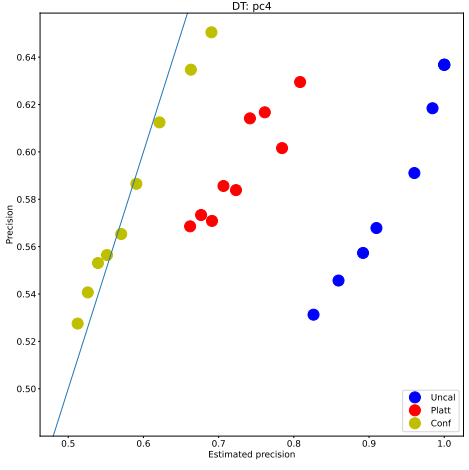
Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated precision and the y-axis as the actual precision obtained.

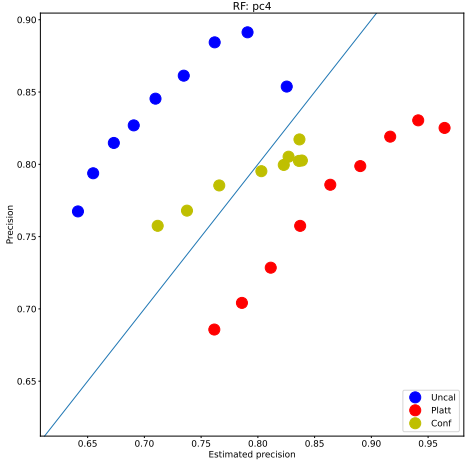




Example: PC4 data set (precision)

Each dot represents a rejection level $\in \{0.1, 0.2, \ldots 0.9\}$, with x-axis as the estimated precision and the y-axis as the actual precision obtained.





Conclusions

6

Conclusions

- We have in this paper demonstrated how conformal classification can be used to produce **perfectly calibrated** classifiers with reject option.
- The empirical evaluation, using ten publicly available data sets, showed that the suggested method produced **very exact** accuracy and precision estimates, for all rejection levels investigated.
- A direct comparison with probabilistic predictors clearly demonstrates the advantage of the conformal approach. Even when calibrating the probabilistic predictors using Platt scaling, the resulting estimations were **outperformed** by the conformal classifiers, in particular for precision.
- Specifically, only the conformal models showed **no systematic bias** when estimating either accuracy or precision for the different rejection levels and using the two underlying models decision trees and random forests.

Thank you!

I will be happy to answer questions now or offline.

Contact: ulf.johansson@ju.se