Applying Conformal Predictions on Public BioAssay Data

Paolo Toccaceli, Ilia Nouretdinov, Alex Gammerman

Computer Learning Research Centre, Dept. of Computer Science, Royal Holloway, Univ. of London

Conformal Predictors

Conformal Prediction: value & **uncertainty**

> **Validity guarantee**
> Given a confidence level, the CP outputs predictions whose error rate does not exceed the chosen confidence level (almost surely) if the test objects come from the same i.i.d. distribution as the training set

Region prediction: the prediction is a **set** of labels.
Point prediction: a single value, with **confidence** (are there competing alternatives?) and **credibility** (how strong is the evidence?).

Compound Activity Prediction poses two problems for CP:
Data volume: computational demands increase more than linearly
Data imbalance: under-represented class can be affected by disproportionately higher error rate
Inductive CP makes CP computationally feasible on large sets.
Mondrian CP makes the validity guarantee per-label.

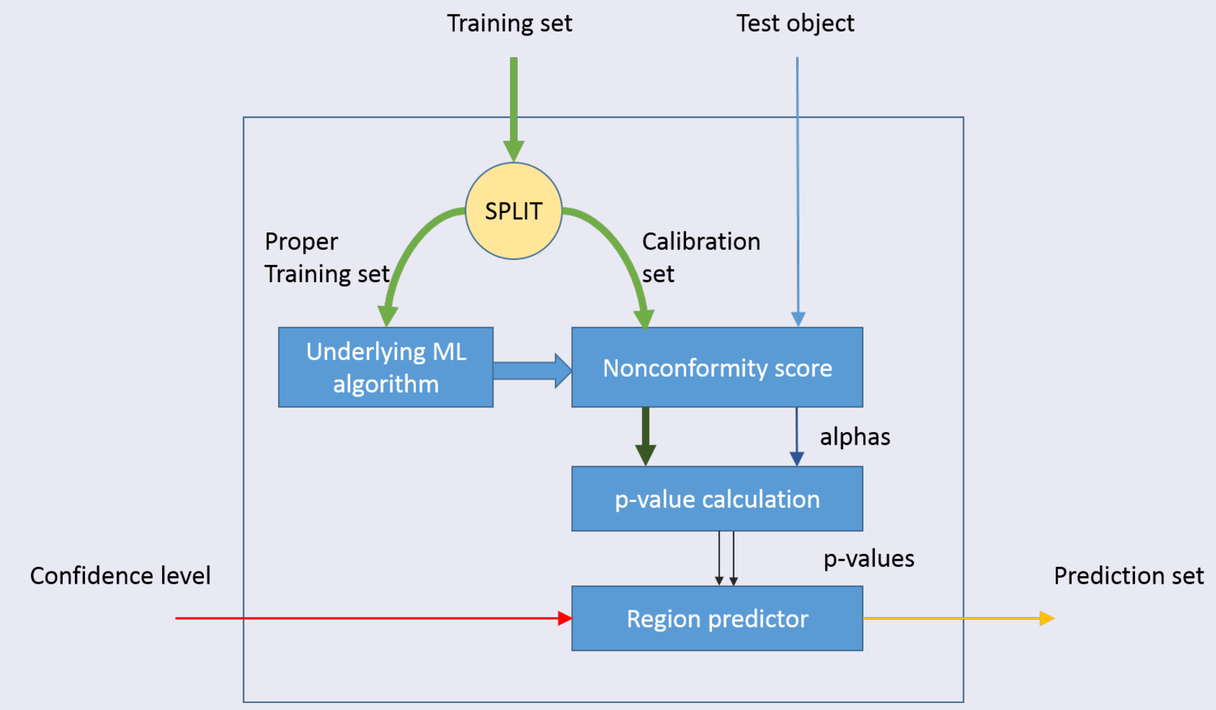

Figure: Inductive Conformal Predictor

Application: PubChem BioAssay AID827

The performance of CP was tested on a data set taken from PubChem BioAssay database on which *signature descriptors* were computed.

Total number of examples	138,287	
Number of features	165,786	High-dimensional
Number of non-zero entries	7,711,571	
Density of the data set	0.00034	Highly sparse
Active compounds	1,658	Highly imbalanced (1.2%)
Inactive compounds	136,629	
Unique set of signatures	137,901	Low degeneracy

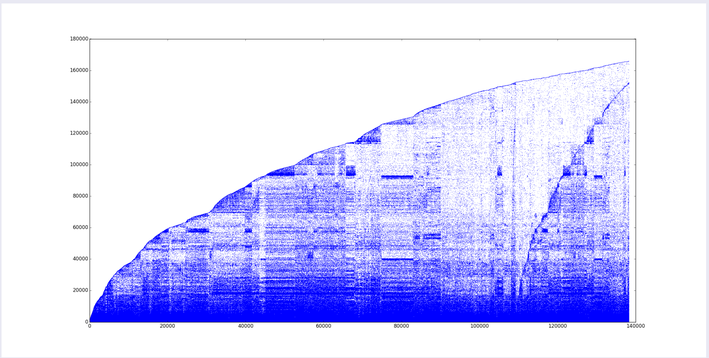

Figure: Sparseness of the data set.

Methodology and tools

Nearest Neighbour, Multinomial Naive Bayes and SVM were used as underlying. Each test consisted of a batch of 20 repetitions with randomly drawn test set, which each batch having the same PRNG seed initialization. Implementation was in Python Jupyter, using `sklearn` for Machine Learning algorithms and `ipyparallel` for parallelization. Execution was performed on the Salomon IT4I supercomputer (Ostrava, Czech Republic) as part of the ExCAPE H2020 European project.

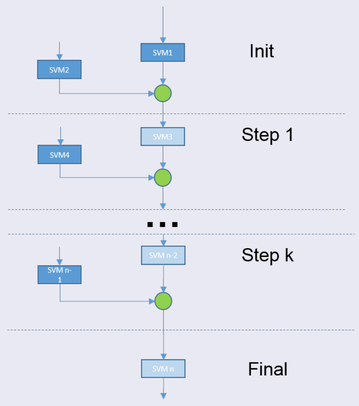

Figure: Our variant of Cascade SVM

The**Tanimoto** similarity was used as one of the distances for NN and one of the kernels for SVM and was implemented in Cython for speed. Training the SVM with 100+k examples was performed using a form of **Cascade SVM** with distributed calculation of the Gram matrix (kernel) and of the decision function. The application of the SVM to this highly-imbalanced set required **per-class weighting** in order to achieve even error distribution over Actives and Inactives.

Results

Underlying	Active pred Active	Inactive pred Active	Inactive pred Inactive	Active pred Inactive	Empty pred	Uncertain
Linear SVM	34.20	99.00	591.85	1.2	0	9273.75
Rbf SVM	47.20	101.80	1126.75	1.8	0	8722.45
Tanimoto SVM	48.45	97.65	986.85	0.8	0	8866.25
TanimotoRbf SVM	53.25	104.35	1202.10	0.6	0	8639.70
M Naive Bayes	38.20	104.30	183.30	1.10	0	9673.10
3NN	43.95	100.55	361.55	0.80	0	9493.15

The table on the left shows the performance of the region predictor at significance $\epsilon = 0.01$, i.e. 99% confidence.

Test set size	10,000
Proper training set size	100,000
Calibration set size	28,287

Conclusions

While the prevalence of Actives in the data set is just 1.2%, 34% of the compounds predicted as active by Inductive Mondrian Conformal Prediction were actually Active. This equates to an **"enrichment" of** $\approx 28\times$ compared to random sampling from the original data set. At the same time, the Recall was $\approx 44\%$ (ratio of Actives in the prediction to total Actives in the test set) .
Conformal Predictors provide the user the ability to select a different trade-off between Precision and Recall by changing the confidence level.

References & Acknowledgements

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.
Alexander Gammerman and Vladimir Vovk. Hedging Predictions in Machine Learning. The Computer Journal, Volume 50, Issue 2, pp. 151-163, 2007.
Vapnik et al., Parallel Support Vector Machines: The Cascade SVM, in *Adv. in Neural Information Processing Systems*, vol. 17, pp. 521–528, MIT Press, 2005
The authors acknowledge Lars Carlsson for providing the data set and the ExCAPE H2020 project under which the study was funded.