ExCAPE WP1. Probabilistic prediction

Paolo Toccaceli*, Ilia Nouretdinov*, Zhiyuan Luo*
Vladimir Vovk*, Lars Carlsson** and Alex Gammerman*

*Royal Holloway, University of London
**AstraZeneca, Sweden

1 Opening remarks

This memo documents the activity of the team working on WP1.4. The objectives of the work package are reported in the box below.

WP1.4 Confidence estimation and feature significance (RHUL and AZ)

- 1.4.1.Optimization of the transductive and inductive conformal prediction framework (CP) for imbalanced and big data with online machine learning [RHUL].
- 1.4.2. Estimation of the feature significance using CP approaches and methods developed in 1.2. [RHUL]; Addressed challenges: confidence estimation.
- 1.4.3. Integration of the CP with methods developed in 1.2 [RHUL, AU, UL].
- 1.4.4. Platt scaling for probabilistic confidence estimation for supervised algorithms developed in Task 1.2. [UL]. Addressed challenges: confidence estimation.

2 Introduction

The objective of this subpackage is to complement the bare prediction of bioactivity with an estimate of its uncertainty. The previous Report [5] described Conformal Prediction and discussed some results of its application to BioAssay data. This Report introduces Multi-probabilistic prediction, also referred to as Venn prediction.

3 Probabilistic Prediction

In this section, we present the main concepts and their rationale. The reader is referred to the many publications for all the details — see, for example, [2].

We also assume that the reader is familiar with the general framework of classification, i.e.

- a *training set* made of ℓ *examples* $(x_i, y_i) \in \mathbf{Z}$, where x_i is an *object* generally represented as a vector of d *attributes* and y_i is a *label* taking a finite number of values, indicating for instance the class to which the example belongs. For simplicity, we'll restrict ourselves to the case of binary classification.
- a *test set* made of objects $x_{\ell+1}, x_{\ell+2}, \ldots$, whose labels we are asked to predict

The approach followed by the Venn Prediction is entirely non-parametric.

At the heart of Venn prediction, there is a notion of taxonomy (note: this is a concept that is completely unrelated to that of taxonomy discussed in the context of Mondrian Conformal Prediction).

Venn taxonomy: A measurable function that assigns to each $n \in \{2, 3, \ldots\}$ and each sequence $(z_1, \ldots z_n) \in \mathbf{Z}^n$ an equivalence relation $\sim$ on $\{1, \ldots, n\}$ which is equivariant in the sense that, for each n and each permutation π of $\{1, \ldots, n\}$,

$$(i \sim j \mid z_1, \ldots z_n) \Rightarrow \big(\pi(i) \sim \pi(j) \mid z_{\pi(1)}, \ldots, z_{\pi(n)}\big)$$

The Venn taxonomy induces equivalence classes which we can define as:

$$A\,(j \mid z_1, \ldots z_n) := \{i \in \{1, \ldots, n\} \mid (i \sim j \mid z_1, \ldots, z_n)\}$$

Venn predictor: Given a training sequence $(z_1, \ldots, z_\ell)$ and a test object $x_{\ell+1}$, the Venn predictor associated with a given Venn taxonomy outputs the pair (p_0, p_1) as its prediction for $x_{\ell+1}$'s label, where

$$p_y = \frac{|\{i \in A(\ell+1 \mid z_1, \ldots, z_\ell, (x_{\ell+1}, y)) \mid y_i = 1\}|}{|A(\ell+1 \mid z_1, \ldots, z_\ell, (x_{\ell+1}, y)|}$$

for $y \in \{0, 1\}$

Intuitively, the equivalence classes of the taxonomy group objects that we consider sufficiently similar for the purposes of prediction. For instance, they could correspond to intervals of the value of the decision function, when using an SVM.

p_y is the empirical probability distribution of the label, calculated as a fraction of the examples, in the same category as the completion $(x_{\ell+1}, y)$, that have label $y_i = 1$.

Both p_0 and p_1 express the probability that the label y of the test example be 1, but under two different hypotheses, namely $y = 0$ and $y = 1$. It is the case that $p_0 \leq p_1$;

informally, one can assume that the actual probability lies "almost surely" between those two values.

The approach described above can be viewed as *transductive* in the sense that we recompute the probability distribution from scratch for every test object. In the SVM example we introduced earlier, one would add the completion $(x_{\ell+1}, y)$ to the training set and retrain the SVM for each of the two possible values of the test object label and for each test object.

Of course this is computationally infeasible. However, similarly to what was discussed for Conformal Prediction, it is possible to have an *inductive* mode of operation, at the cost of losing some theoretical guarantees. We'll discuss it in the context of Venn-ABERS predictors [3].

3.1 Validity of Venn predictors

Let's say that a random variable P taking values in $[0, 1]$ is *perfectly calibrated* for a random variable Y taking values in $\{0, 1\}$ if

$$\mathbb{E}(Y \mid P) = P \quad \text{almost surely}$$

Intuitively, P is the prediction made by a probabilistic predictor for Y, and perfect calibration means that the probabilistic predictor gets the probabilities right, at least on average, for each value of the prediction. We then say that the probabilistic predictor is *valid*.

Unfortunately, it can be proved that it is not possible to make point probabilistic predictions that are also valid.

However, it can be proved that if training set and test set are IID, the Venn predictor is well calibrated in the sense that there exists a choice of the output probabilities that is well calibrated (although we do not know how to choose the right probability for each test object).

4 Venn-ABERS Predictors

The Venn-ABERS predictors apply the established technique of Isotonic Regression (IR) within the framework of Venn prediction. There are two advantages accruing from this approach: one theoretical, as we gain the validity guarantees of Venn predictors, and one practical, as the notorious IR tendency to overfit is lessened.

Venn-ABERS predictors can be applied onto classification algorithms that use a form of scoring, that is, a continuous value that is then thresholded to obtain the prediction.

Let $s_0(x)$ be the scoring function for $(z_1, z_2, \ldots, z_\ell, (x, 0))$, $s_1(x)$ be the scoring function for $(z_1, z_2, \ldots, z_\ell, (x, 1))$, $g_0(x)$ be the isotonic calibrator for

$$((s_0(x_1), y_1), (s_0(x_2), y_2), \ldots, (s_0(x_\ell), y_\ell), (s_0(x), 0))$$

and $g_1(x)$

$$((s_1(x_1), y_1), (s_1(x_2), y_2), \ldots, (s_1(x_\ell), y_\ell), (s_1(x), 1))$$

The *isotonic calibrator* g for $((s(x_1), y_1), (s(x_2), y_2), \ldots, (s(x_\ell), y_\ell))$ is the increasing function on $s(x_1), s(x_2), \ldots, s(x_\ell)$ that maximizes the likelihood

$$\prod_{i=1,2,\ldots,\ell} p_i$$

where:

$$p_i = \begin{cases} g(s(x_i)) & \text{if } y_i = 1 \\ 1 - g(s(x_i)) & \text{if } y_i = 0 \end{cases}$$

The Venn-ABERS predictor outputs a multi-probabilistic prediction (p_0, p_1), where $p_0 := g_0(s_0(x))$ and $p_1 := g_1(s_1(x))$

4.1 Inductive VAP

The definition of VAP in the previous paragraphs requires that the IR be recalculated for each possible label value and for each test object. The resulting computational cost would become unaffordable even for relatively small data sets.

It is possible to reduce significantly the computational cost by adopting an inductive approach, similar in a broad sense to the procedure described for Inductive Conformal Predictors in the previous report.

The overall training set is split into a proper training set, which is used to train the underlying machine learning algorithm which produces scores, and a calibration set, which is used to "train" the isotonic calibrator which transforms scores into probability estimates.

The training of the underlying machine learning algorithm is performed only once, but the isotonic regression is still recalculated (ex novo) for each test object and for each possible value of the label.

While this scheme allows us to reduce significantly the computational cost of VAP, there is no claim that it enjoys the same theoretical validity guarantees that "proper" Venn predictors have. However, in practice, this does not appear to be a significant shortcoming.

4.2 Cross IVAP

A potential disadvantage of Inductive VAP is that, by separating a calibration set from the overall training set, it reduces the size of the set used for training the underlying algorithm. It may be of benefit to use a scheme inspired by cross-validation in which the overall training set is split into K sets; IVAP is then applied K times, each time using one of these sets K as calibration set and the rest as proper training set. The resulting K IVAPs can then be combined in way that will become clearer in the next section.

4.3 Making probability predictions out of multi-probability ones

While there is a value in having multi-probability predictions, it is generally more intuitive to deal with a point probability prediction. One way to construct a probability prediction from a multi-probability prediction is to derive the value that minimizes the regret[1] under a given loss function.

It is possible to prove that under log-loss function the minimax probabilistic prediction is:

$$p = \frac{p_1}{1 - p_0 + p_1}$$

Note that p is calculated as if $1 - p_0$ and p_1 were the unnormalized probabilities of $y = 0$ and $y = 1$, respectively.

If, instead of the log-loss function, we consider the square loss function (also known as Brier loss), the minimax probabilistic prediction is:

$$p = \bar{p} + (p_1 - p_0)\left(\frac{1}{2} - \bar{p}\right)$$

where $\bar{p} := (p_0 + p_1)/2$.

In the case of Cross VAP, the K multi-probabilistic predictions can be combined as in:

$$p = \frac{\mathrm{GM}\,(p_1)}{\mathrm{GM}\,(1 - p_0) + \mathrm{GM}\,(p_1)}$$

where GM() stands for geometric mean over the K values.

4.4 Fast Venn-ABERS

While it reduces the computational load significantly (compared to the original "transductive" formulation), Inductive Venn-ABERS prediction still remains too complex to scale to large data sets, the recalculation of the Isotonic Regression for every label value and for every test object being the main limiting factor. From a detailed analysis of the algorithm used to compute the IR and of the particular way in which it used in the Venn-ABERS application, it was possible to come up with a new algorithm that produces exactly the same results, but requires one initial *pre-calculation* ($O(\ell \log \ell)$) and then requires for the actual *evaluation* of the probability of each test object only a very efficient look-up ($O(\log \ell)$). The full details can be found in [4]

An illustration of the Isotonic Calibrator on synthetic data sets is provided in Fig. 1.

Tab. 1 gives an indication of the performance, on synthetic data sets, of a pure Python/Numpy implementation running on one core of our departmental server. (Implementing it in C/C++ is likely to reduce the quoted times by factor of 10 or more).

This enabled us to apply Venn-ABERS prediction to large data sets (hundreds of thousands of examples) without particular requirements in terms of computational resources.

[1] By regret, we denote the loss suffered by making a given choice instead of the best available option.

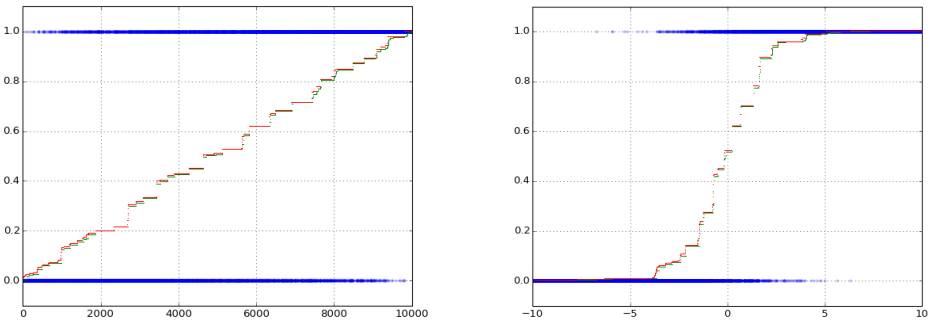

Figure 1: Multi-probabilistic predictions for two synthetic data sets. The x axis represents the scores. The blue dots represent the data, which can take values 0 or 1. The data sets were created so that the desired calibrator is a linear function in the left pane and a sigmoid in the right. This was by generating random variates so that their probability of being 1 was proportional to the score (left pane) or to a sigmoid of the score (right pane). The Venn-ABERS predictors reconstruct the desired calibrator within an error inherent in the statistical fluctuation of the sample.

Table 1: Execution times of the pre-calculation and evaluation of the Fast Venn-ABERS IR. The number of examples quoted refers to the size of the calibration set as well as the test set.

Number of examples	CPU time for Pre-calculation on calibration set	CPU time for Evaluation on test set
10,000	3.62s	<0.01s
100,000	15.7s	0.01s
1,000,000	153s	0.13s

5 Application of Venn-ABERS Prediction to BioAssay Data

We applied VAP to the same `827.svm` data set used in the previous report for the CP studies.

To recap the salient points, each tested compound is described by a variable number of *signature descriptors* [1] derived for us by AZ from the chemical structure of the compounds itself. Each signature corresponds to the number of occurrences of a given labelled subgraph in the molecule graph. The resulting data set can be viewed as a relatively sparse matrix of attributes (the signatures on the columns) and examples (the compounds on the rows).

Total number of examples	=	138,287
Number of features	=	165,786
Number of non-zero entries	=	7,711,571
Density of the data set	=	0.00034
Active compounds	=	1,658
Inactive compounds	=	136,629
Unique set of signatures	=	137,901

We reused the same values of the decision function calculated in that context for the calibration set and test set for the Inductive Conformal Prediction. As a side note, Fig. 2 shows the distribution of such values, with two separate histograms one for Active examples and the other Inactive Examples. The distribution for Active examples looks distinctly bimodal.

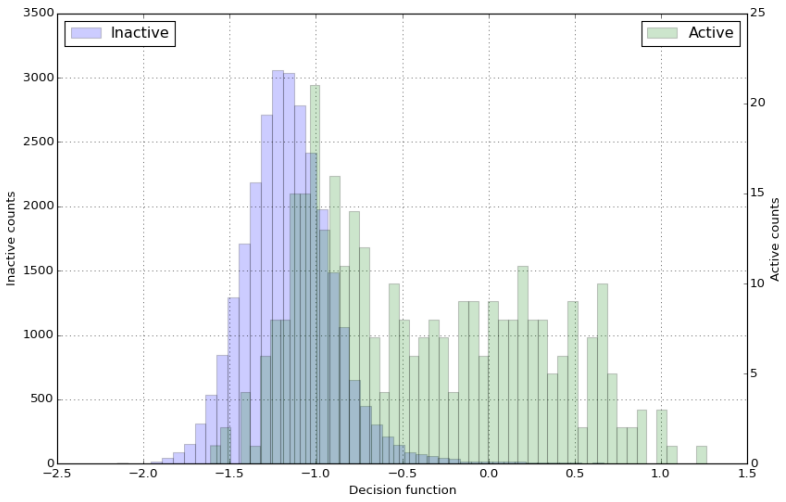

Figure 2: Distribution of Decision Function values on the calibration data set. The numbers on the left and right y axes refer to the counts per bin. Note the different y scales.

The calibration set has 28,000 examples and the test set comprises 10,000 objects. Running the Fast Venn-ABERS predictors on these sets took a total time 4.84s on one core of the departmental server (of course, this excludes the time to calculate the scores, given that we reused them, as explained above).

One illustration of the multi-probabilistic results is provided in Fig. 3 which shows the cumulative sums of p_0, of p_1, and of the label as we sweep the test set.

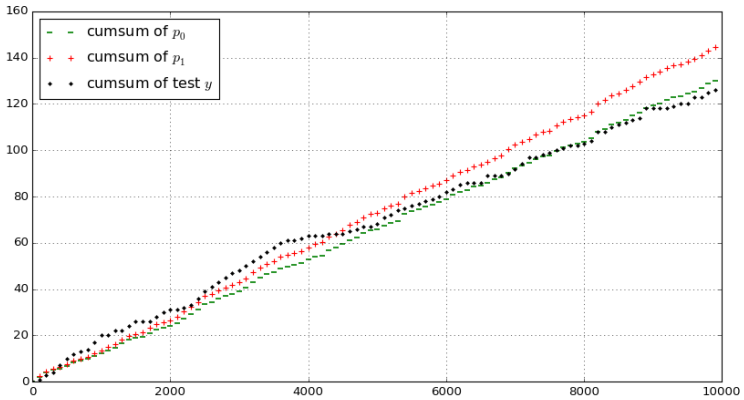

Figure 3: Data set 827: Cumulative sums of p_0,p_1,and of the label, i.e. $\sum_{i\leq x} p_{0,i}$ (green), $\sum_{i\leq x} p_{1,i}$ (red), $\sum_{i\leq x} y_i$ (black)

.

Fig. 4 has a plot of the "calibrators" $g_0(s)$ and $g_1(s)$ between the min and max of the decision function on the calibration set. The interesting feature is the widening gap between $g_0(s)$ and $g_1(s)$ for $s > 0.5$ indicative of uncertainty on the probability estimate.

5.1 Comparison with Platt's scaling

We set out to compare the relative merits of Platt's scaling and Venn-ABERS predictors. We use the implementation of Platt's scaling available as part of the Python `sklearn` package.

Fig. 5 shows the "calibrators" arising from the two methods (for one of 20 runs).

In order to have a more meaningful comparison, we computed the log-loss with either method on the test set of each of the 20 runs. The resulting boxplots are shown in Fig. 6. The median is the same (VA is better by the tiniest of margins) and, if anything, Platt's method exhibits a bit more variance.

Given the difference between the two calibrators apparent in Fig. 5, one might have expected a difference also in the log-loss.

Imbalance plays a role in this apparent paradox. For values of the SVM Decision Function (DF) less than -0.7, there is very little difference between the two calibrators.

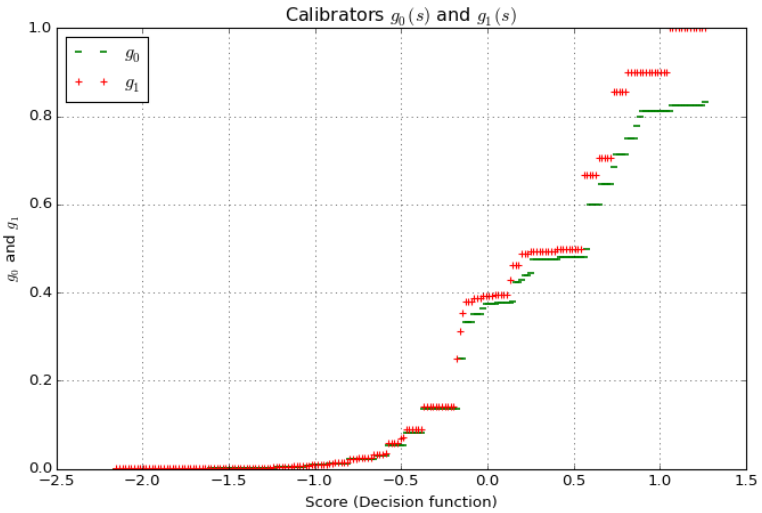

Figure 4: Data set 827: Calibrators $g_0(s)$ and $g_1(s)$

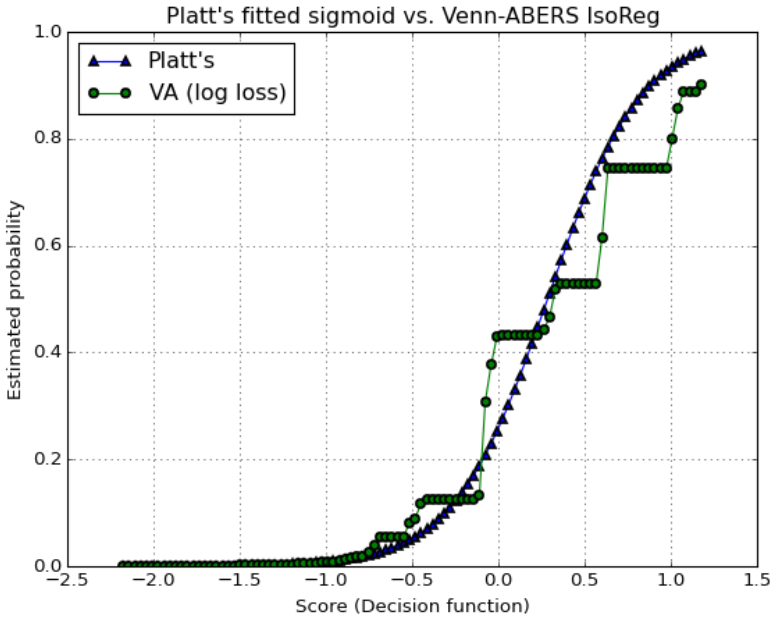

Figure 5: Data set 827: Comparison of calibrators

.

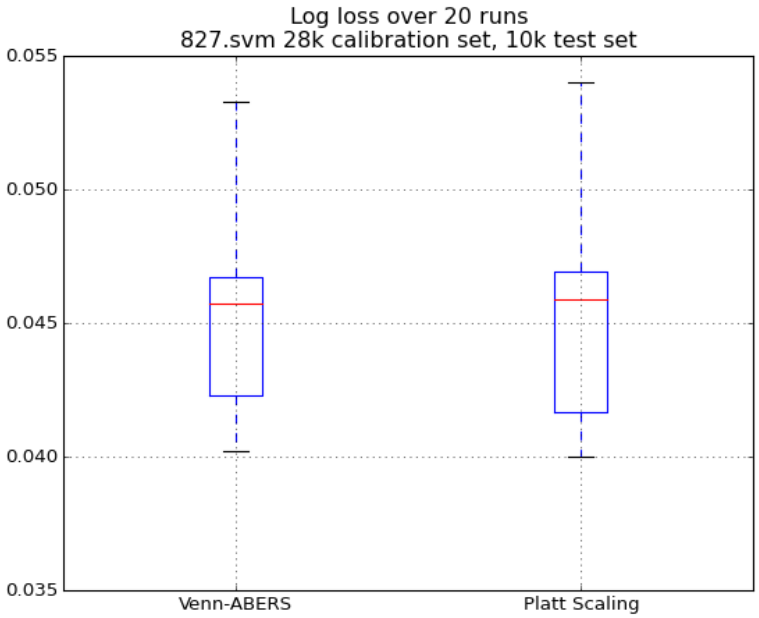

Figure 6: Data set 827: Log-loss on test sets, over 20 runs

.

Consider that $\approx 95\%$ of the test samples happen to have a DF< -0.7. For those test samples the difference between Platt's scaling probabilities and Venn-ABERS probabilities is going to be negligible.

It is questionable if the loss is "symmetrical" in our application, i.e. if the errors in predicting high probability have the same consequences and should be attributed the same cost as the errors of the same entity for low probabilities.

One way to have an asymmetric log-loss is to assign different weights to the loss on Actives and on Inactives. Unfortunately, this would make the loss function no longer *proper*[2].

$$\text{AsymmLogLoss} = -\frac{1}{N}\sum_{i=1}^{N} C_{act} y_i \log p_i + C_{inact}(1-y_i)\log(1-p_i)$$

If one sets $C_{act} = 1$ and $C_{inact} = 0$, thereby disregarding any loss on Inactive test examples, the effect of the difference between the two calibrators for higher DF values becomes now visible as shown in Fig. 7 (note that we are still averaging across all the

[2]Limiting ourselves to the case of binary classification, a loss function $L(q, y)$ is proper [6][7] when the minimum value of the expected loss is attained when the predicted probability is equal to the actual probability. In more formal terms, when $\underset{q}{\operatorname{argmin}}\, \mathbb{E}_{y\sim B_p}[L(q, y)] = p$.

test examples, which are 99% Inactive).

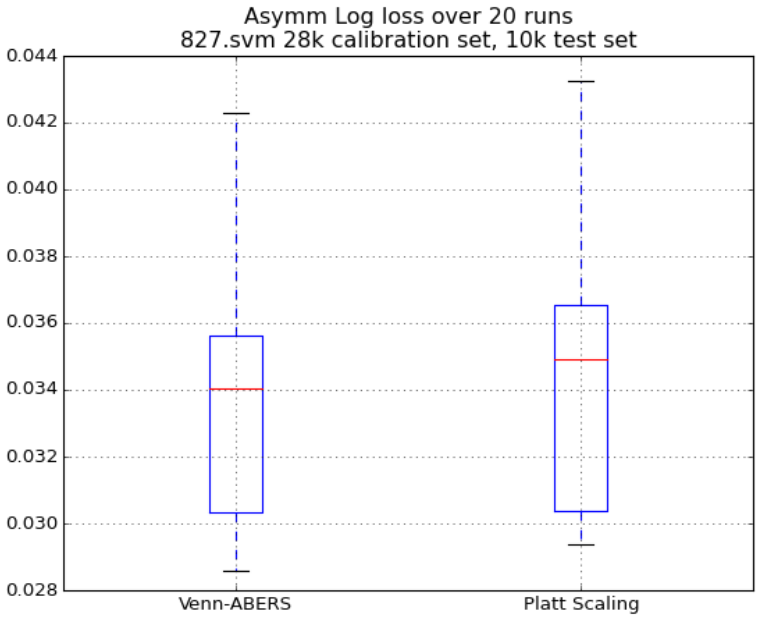

Figure 7: Data set 827: Asymmetric Log-loss on test sets, over 20 runs

.

6 Comparison between Conformal and Probabilistic results

Both Conformal and Probabilistic Predictors allow the user to rank the compounds for the purposes of screening. The two rankings are based on conceptually different ideas. The p-value of Conformal Prediction expresses the probability of encountering an Active compound that would appear as or more nonconform (among Active compounds) than the test object. The value output by the Probabilistic Prediction expresses the probability that the compound is Active (in the sense that if one keeps taking compounds for which the predicted probability is 0.8, then one will find in the long term that 80% of such compounds are actually active).

Tab. 2 shows the 20 compounds (out of a test set of 10,000) that are ranked highest by Conformal Prediction and by Venn-ABERS Prediction. The compounds are identified by an ordinal that has no specific meaning outside of the specific data set we were supplied. The single probability was obtained from the multi-probabilistic prediction using the method discussed earlier for log-loss regret minimization.

Although the ranking may appear quite different at first glance, it is remarkably similar. The similarity is somewhat obscured by the fact that the p-values appear to have a finer granularity than the probabilities (there are more compounds sharing probabilities than sharing p-values). The rankings appear very similar if we consider that the probability ranking is arbitrary when the compounds have the same probability.

Table 2: The 20 candidate compounds (out of a test set of 10,000) that are ranked highest, according to Conformal Prediction and to Venn-ABERS Prediction. The asterisk denotes those compounds that are actually Active.

Ranking	Best compounds by p-value	Best compounds by probability	p-value	Prob
0	*108198	*108198	0.868	0.800
1	103948	129143	0.798	0.697
2	62772	62772	0.774	0.697
3	129143	103948	0.716	0.697
4	*138051	*138051	0.663	0.695
5	*108632	*108632	0.663	0.695
6	*107957	*108877	0.584	0.683
7	*108920	*108920	0.584	0.683
8	*108877	*107957	0.584	0.651
9	*108783	*108783	0.578	0.651
10	5413	5413	0.557	0.647
11	*138177	54806	0.551	0.611
12	4334	16925	0.537	0.611
13	71538	*108032	0.513	0.611
14	54806	*108026	0.513	0.611
15	16925	4334	0.493	0.611
16	*108584	71538	0.490	0.611
17	*108026	*138177	0.490	0.611
18	*107943	*107943	0.478	0.611
19	*108032	*108584	0.475	0.611

Fig. 8 illustrates another facet of the relationship between p-values and probabilities. A visual representation of their similarity can be seen in terms of how close the dots, each representing one of the 10,000 compounds in the test set, are to the (0,0),(1,1) line.

7 An example on the official data: Data Set 3717

A new data set for ExCAPE was released in April. Out of this set, covering a large number of biological targets, we obtained a single-target data set and applied Conformal

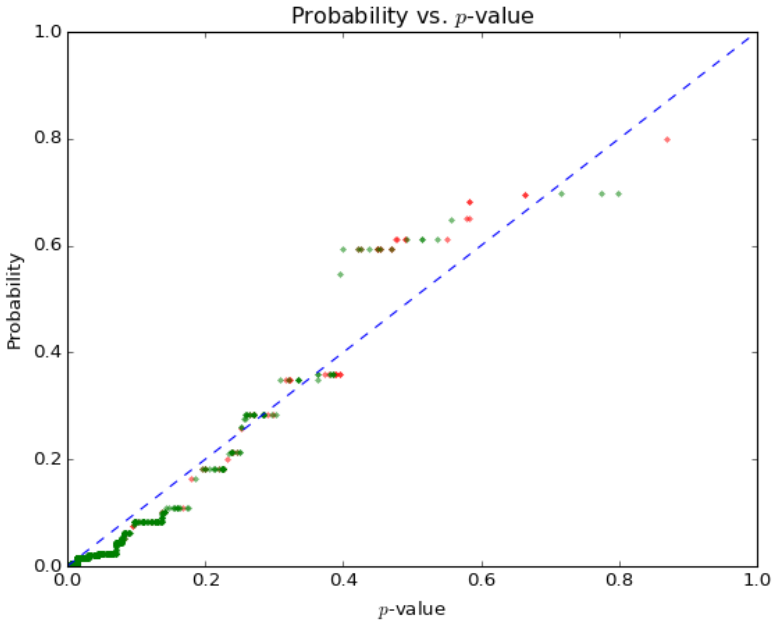

Figure 8: Data set 827: Each of 10,000 test objects is represented as a dot. Red dots are Active compounds, green dots are Inactive compounds.

.

Prediction and Probabilistic Prediction to it.

The target for the data set in hand is identified by Gene ID 3717 and the main characteristics of the data set are summarised below:

Total number of examples	=	215,129
Number of features	=	219,772
Number of non-zero entries	=	11,998,132
Density of the data set	=	0.025%
Active compounds	=	2,077
Inactive compounds	=	213,052

This data set appears to lend itself very well to prediction. The distributions of the values of the Cascade SVM DF for Active and Inactive examples in the calibration set are illustrated in Figure 9. The Cascade SVM achieved an AUC of 98.3%, a Precision of 93.7% and a Recall of 81.7% (averaged over 20 repetitions of Cascade SVM using Tanimoto+RBF kernel with randomly drawn training sets (170,000 examples) and test sets (10,000 objects)). The confusion matrix is reported in Table 3.

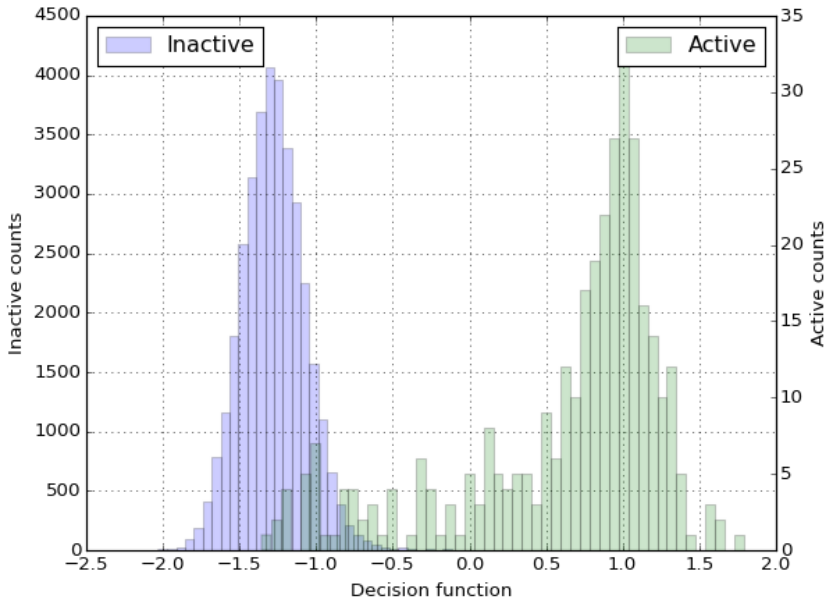

Figure 9: Data set 3717: One example of the empirical distribution (histogram) of Decision Function values for Actives (green) and Inactives (pale blue). Note the clear separation between the two classes.

Figures 12, 13, 14 show for data set 3717 the same information that was shown in Figure 5, 6, 7 for data set 827.

Table 3: Confusion matrix averaged over 20 trainings of Cascade SVM with randomly drawn training sets (170,000 examples) and test sets (10,000 objects)

	Predicted Inactive	Predicted Active
Actual Inactive	9897.10	5.35
Actual Active	17.85	79.70

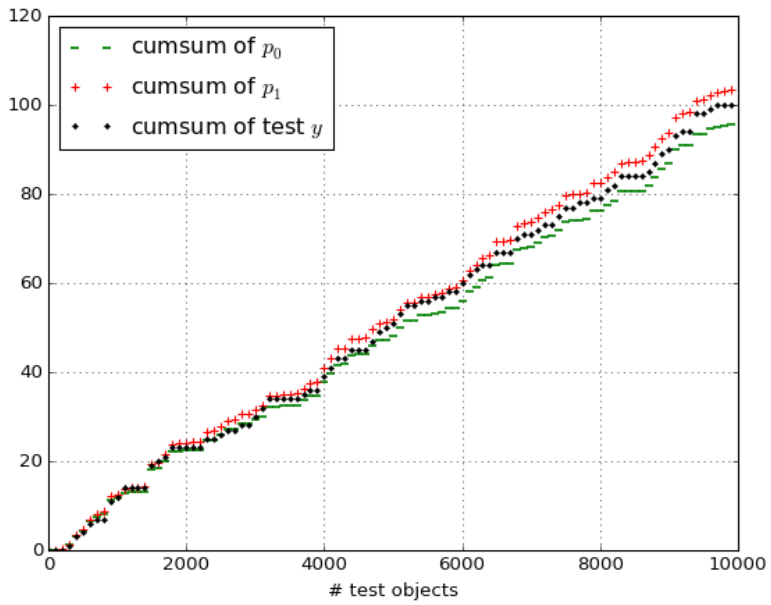

Figure 10: Data set 3717: Example of cumulative sums of p_0,p_1,and of the label, i.e. $\sum_{i \leq x} p_{0,i}$ (green), $\sum_{i \leq x} p_{1,i}$ (red), $\sum_{i \leq x} y_i$ (black)

.

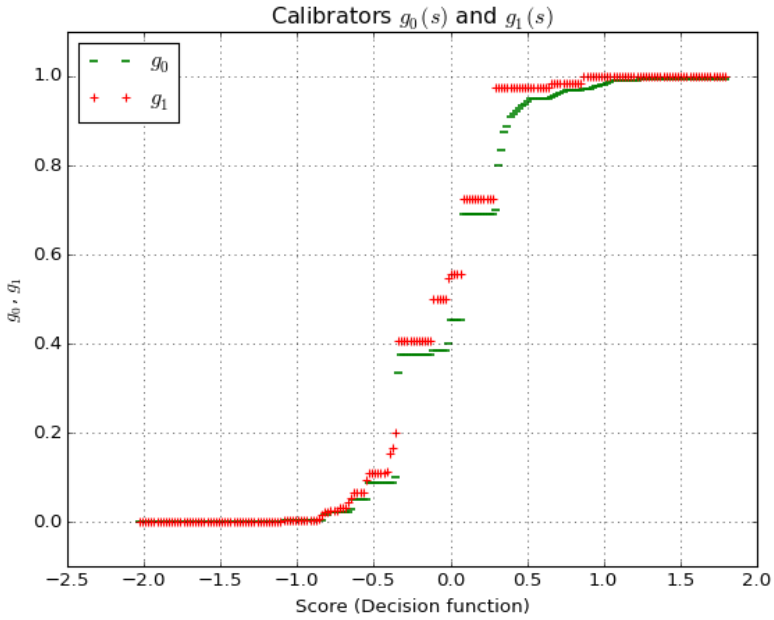

Figure 11: Calibrators $g_0(s)$ and $g_1(s)$

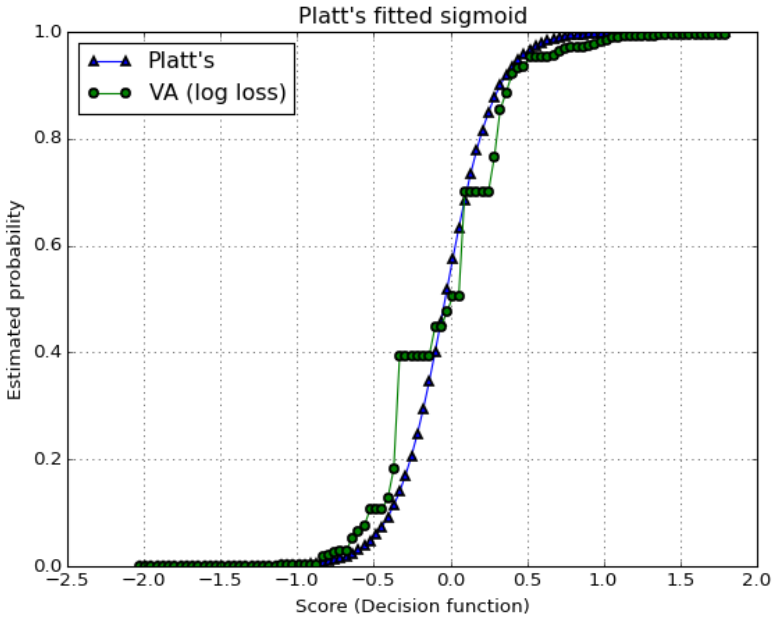

Figure 12: Data set 3717: Platt's and Venn-ABERS calibrators

.

The comments for data set 3717 are similar to those for 827. Perhaps the difference between Platt's and Venn-ABERS calibrators is even less significant, but there still is a tiny advantage in log loss terms for Venn-ABERS. Also for data set 3717, the log loss for Venn-ABERS appears to exhibit lower variance.

The list of the top 20 candidate test compounds in Table 4 by p-value and by probability confirms the findings already seen in the case of data set 827. The ranking resulting from the Conformal Predictor has a finer granularity than that resulting from Venn-ABERS (in the sense that more compounds share the same probability than the same p-value). The rankings are identical if one disregards the arbitrary ordering in case of ties.

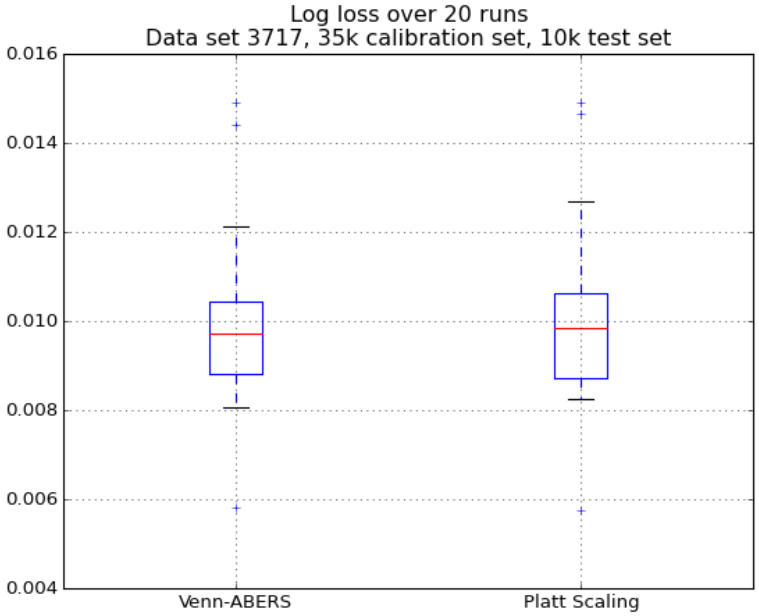

Figure 13: Data set 3717: Log-loss on test sets, over 20 runs

.

8 Summary and Conclusions

In this report we introduced Probabilistic Prediction and highlighted its conceptual differences with Conformal Prediction. We presented the notion of validity for a probabilistic predictor and introduced Venn Predictors, multi-probabilistic predictors for which it is possible to prove the validity property. We then turned to Venn-ABERS predictors, which are Venn Predictors that calibrate classification scores into probabilities. Finally, we applied a fast method of computing Venn-ABERS probabilities to a bioassay data set and compared and contrasted its results with Platt's scaling and Conformal Predictors.

Table 4: Data set 3717: The 20 candidate compounds (out of a test set of 10,000) that are ranked highest, according to Conformal Prediction and to Venn-ABERS Prediction. The asterisk denotes those compounds that are actually Active.

Ranking	Best compounds by p-value	Best compounds by probability	p-value	Prob
0	*214891	*214694	1.000	0.994
1	*213784	*213784	1.000	0.994
2	*214318	*214891	1.000	0.994
3	*214694	*214318	1.000	0.994
4	*214944	*214944	0.997	0.993
5	*495	*214714	0.997	0.993
6	*214741	*213852	0.997	0.993
7	*214714	*213853	0.997	0.993
8	*213852	*214741	0.997	0.993
9	*213853	*495	0.997	0.993
10	*214055	*214789	0.994	0.993
11	*213601	*213601	0.994	0.993
12	*214789	*214055	0.994	0.993
13	*501	*309	0.991	0.993
14	*214808	*501	0.991	0.993
15	*309	*214511	0.991	0.993
16	*213851	*213851	0.991	0.992
17	*214511	*214808	0.991	0.992
18	*214811	*214811	0.988	0.992
19	*213899	*214342	0.982	0.992

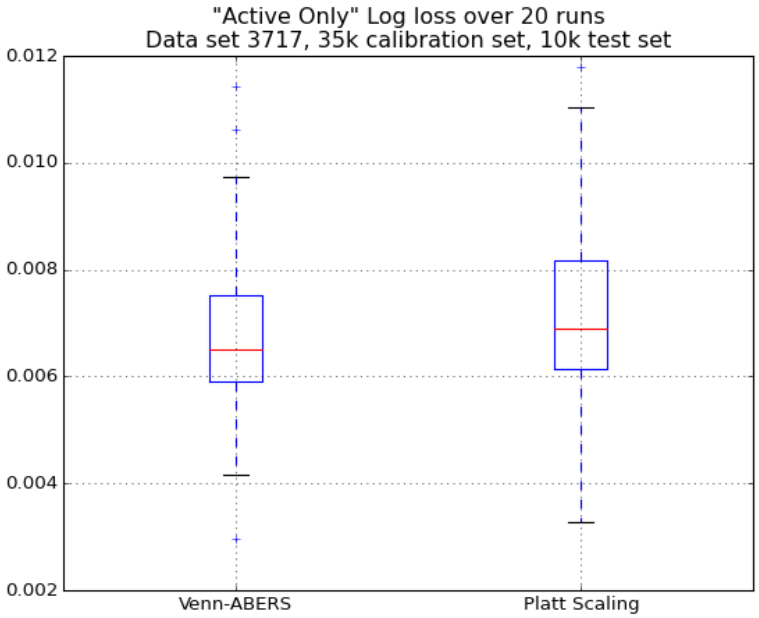

Figure 14: Data set 3717: Log-loss for Active examples only on test sets, over 20 runs.

Venn-ABERS predictors can improve on Platt's scaling, in particular when the functional dependency of probabilities to scores departs significantly from a sigmoid. Their computational cost is competitive with Platt's scaling and perfectly affordable even on large data set sizes.

Conformal Predictors and Venn-ABERS Predictors appear to produce a very similar ordering of compounds when ranking them to identify the more promising candidates. Probabilistic Prediction offers a more theoretically sound basis for merging predictions from different classifiers (using different kernels or learning on separate subsets of the training data).

References

[1] Jean-Loup Faulon, Jr. Donald P. Visco, and Ramdas S. Pophale.
The signature molecular descriptor. 1. using extended valence sequences in qsar and qspr studies.
Journal of Chemical Information and Computer Sciences, 43(3):707–720, 2003.
PMID: 12767129.

[2] Vladimir Vovk, Alex Gammerman, and Glenn Shafer.
Algorithmic Learning in a Random World.
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[3] Vladimir Vovk and Ivan Petej.
Venn-Abers predictors.
In Proceedings of 30th Conference on Uncertainty in Artificial Intelligence, 2014.
`http://alrw.net/articles/07.pdf`

[4] Vladimir Vovk, Ivan Petej, and Valentina Fedorova.
Large-scale probabilistic prediction with and without validity guarantees.
In NIPS, pages 892-900. 2015 `http://alrw.net/articles/13.pdf`

[5] Paolo Toccaceli, Ilia Nouretdinov, and Alex Gammerman.
ExCAPE WP1. Conformal Predictors.
`http://www.clrc.rhul.ac.uk/projects/ExCAPE/Report_wp1_dec05a.pdf`

[6] Brier, G. W.
Verification of forecasts expressed in terms of probability.
Monthly Weather Rev. (1950), 78, 13.

[7] A. Philip Dawid.
Probability forecasting.
In Samuel Kotz, N. Balakrishnan, Campbell B. Read, Brani Vidakovic, and Norman L. Johnson, editors, Encyclopedia of Statistical Sciences, volume 10, pages 6445-6452. Wiley, Hoboken, NJ, second edition, 2006.