Combination of Conformal Predictors for Classification

Paolo Toccaceli PAOLO.TOCCACELI@RHUL.AC.UK
Computer Learning Research Centre
Royal Holloway, University of London
Egham, UK
Alexander Gammerman ALEX@CS.RHUL.AC.UK
Computer Learning Research Centre
Royal Holloway, University of London
Egham, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

The paper presents some possible approaches to the combination of Conformal Predictors in the binary classification case. A first class of methods is based on p-value combination techniques that have been proposed in the context of Statistical Hypothesis Testing; a second class is based on the calibration of p-values into Bayes factors. A few methods from these two classes are applied to a real-world case, namely the chemoinformatics problem of Compound Activity Prediction. Their performance is discussed, showing the different abilities to preserve of validity and improve efficiency. The experiments show that P-value combination, in particular Fisher's method, can be advantageous when ranking compounds by strength of evidence.

Keywords: Conformal Prediction, Confidence Estimation, Chemoinformatics, Non-Conformity Measure

1. Introduction

Conformal Predictors (CP) (Vovk et al., 2005; Gammerman and Vovk, 2007) provide a theoretically sound way to generate predictions with a chosen rate of errors. This property, referred to as *validity*, is of considerable interest in many application domains. CP prescribe the way to generate prediction sets (so the prediction is multi-valued, as opposed to being a single value, as it is generally the case), so that the validity property is guaranteed. It is of course desirable that the prediction sets be as small as possible. A CP that outputs smaller prediction sets than another is said to be more *efficient*. Since validity is guaranteed, the challenge becomes one of improving efficiency. The efficiency of a specific CP depends on the specific Machine Learning algorithm, referred to as the *underlying* algorithm, that the CP is built on. More accurate underlying algorithms result in smaller prediction sets, hence in higher efficiency of the CP.

The objective of this paper is to explore ways to improve Conformal Prediction by some form of *ensembling*. Ensembling appears to be a recurrent theme of winning submissions to Machine Learning contests. In itself, the term ensembling can be taken to designate different specific strategies. For instance, Bagging and Random Forests aggregate multiple potentially overfitting classifiers, whereas Mixture of Experts both foster specialization in the component classifiers and learn which to choose for a given test object. The form of ensembling investigated in this paper differs from either strategy. It differs from Bagging and Random Forests in that it does not explicitly aim at combating overfitting and correlation per se; it differs from Mixture of Experts in that it takes the component classifiers as a given and does not ”encourage” their specialization. The approach investigated in this paper takes its motivation from the intuition that intrinsically different algorithms are going to make idiosyncratic errors in different parts of the data space and with different modalities. The challenge is to find a method of wide applicability that combines the predictions in a synergistic way.

2. Conformal Predictors

This short section recalls succinctly the key facts about Conformal Predictors. For a gentler introduction the reader is referred to (Shafer and Vovk, 2008; Toccaceli et al., 2016). Assuming that the training set is made up of ℓ independent identically distributed examples (iid)[1] (x_i, y_i), if $x_{\ell+1}$ is a test example taken from the same distribution as the training examples, a Conformal Predictor assigns a p-value to a hypothetical completion $(x_{\ell+1}, y_{\ell+1})$, i.e. a hypothetical assignment of a label $y_{\ell+1}$ to the object $x_{\ell+1}$. The definition of p-value in this context relies on the notion of Non-Conformity Measure (NCM). The NCM is a real-valued function $\alpha(z; z_1, \ldots, z_k)$ that expresses how dissimilar an example appears to be with respect to a bag (or multi-set) of examples, assuming they are all iid. A Non-Conformity Measure can be extracted from any machine learning algorithm, although there is no universal method to choose it.

Armed with an NCM, it is possible to compute for any example (x, y) a *p-value* that has the following property: for any chosen $\epsilon \in [0, 1]$, the p-value of test examples (x, y) drawn iid from the same distribution as the training examples are (in the long run) smaller than ϵ with probability at most ϵ.

The idea is then to compute for a test object a p-value of every possible choice of the label.Once the p-values are computed, they can be put to use in one of the following ways:

- Given a significance level ϵ, a *region predictor* outputs for each test object the set of labels (i.e., a region in the label space) such that the actual label is not in the set no more than a fraction ϵ of the times. This is called the *validity* property. It provides a long term guarantee on the number of errors (where “error” is defined as “actual label not in the prediction set”) in the long run. If the prediction set consists of more than one label, the prediction is called *uncertain*, whereas if there are no labels in the prediction set, the prediction is *empty*.

- Alternatively, one can take a *forced* prediction (where the label with the largest p-value is chosen for a given test object), alongside with its *credibility* (the largest p-value) and *confidence* (the complement to 1 of the second largest p-value).

There are two forms of CP: Transductive CP (TCP) and Inductive CP (ICP). TCP is computationally expensive as the computation of the NCM is performed from scratch

[1] in fact, even a weaker requirement of *exchangeability* is sufficient.

for each object. Inductive CP instead requires just one training of the underlying, but it requires that the training data set be split into a proper training set (to train the underlying) and a calibration set (which is used to compute the NCM). Both the Transductive form and the Inductive form of CP are proven to have the validity property.

Finally, the validity property as stated above guarantees an error rate over all possible label values, not on per-label value basis. The latter can be achieved with a variant of CP, called Mondrian CP. The label-conditional validity guarantee of Mondrian CP is particularly relevant when the distribution of the label values is imbalanced.

3. Requirements for CP combination

The study of the problem of combining p-values to obtain a single test for a common hypothesis has a long history, originating very soon after the framework of statistical hypothesis testing was established (Fisher, 1932). A survey can be found in (Loughin, 2004). In its more general form, the problem raised a lot of attention for its application to meta-analyses, where the results of a number independent studies, generally with different sample sizes and different procedures, are combined. The various methods that have been proposed over the years have tried to cater for the different ways in which the evidence manifests itself. In particular, some methods allow for weighting, thereby assigning more importance to some p-values (for instance, in the case of meta-analyses, those corresponding to studies with larger samples sizes). More importantly, each method is associated with a different shape of the rejection region (the portion of the k-dimensional space of the k p-values being combined for which the combined test of significance would reject the null hypothesis under a chosen significance level ϵ). The shape reflects the different way in which evidence of different strength is incorporated into the aggregated p-value. It has been observed that there is no single combination method that outperforms all others in all applications.

The combination of p-values from different Conformal predictors on the same test object is a very special form of the general problem outline above.

A method for the combination of Conformal Predictors should aim to:

- **Preserve validity**: for the output of the combination method to be a Conformal Predictor, this is a necessary property.
- **Improve efficiency**: smaller prediction sets must result from a desirable method of combination.

In practice, one is interested in the two desiderata above if the resulting p-values are to be used to obtain prediction sets. There are domains of application where the p-values can be used in other ways. An example which will be developed further in the sequel is in the context of Drug Discovery: the p-values can be used to rank candidate compounds in terms of the confidence in their activity (or lack of confidence in their inactivity), so that an informed decision can be made as to which candidate compounds to choose for a new batch of screenings.

There are two key observations that apply to p-values computed by Mondrian Inductive Conformal Predictors (MICP):

1. *The p-values from the same Conformal Predictor for the various test objects do not necessarily follow the uniform distribution.* The p-values in Statistical Hypothesis Testing are uniformly distributed by construction if the null hypothesis is true. Similarly, when one examines the MICP p-values for a set of test objects, it is apparent that only those for which the hypothetical label assignment is the correct one are uniformly distributed. The p-values for the objects for which the hypothetical label assignment is incorrect tend to have values towards 0.

2. *The p-values from different Conformal Predictors for the same test object are not independent.* One has to expect that, when testing the same hypothesis with different methods on the same object, the results will exhibit some degree of correlation. In other applications of p-value combination, the issue may be less of a concern. For instance, in meta-analyses of clinical trials, it is arguable that there is less correlation because the trials are not reusing the same patients in the same groups (hopefully). However, the one considered is certainly not the only context in which dependent p-values are encountered and the issue has attracted some attention by statisticians.

4. Methods from Statistical Hypothesis Testing

As outlined in (Loughin, 2004), there are, broadly speaking, two classes of p-value combination methods: quantile methods and order-statistic methods.

Order-statistic methods (Davidov, 2011) are mentioned here for completeness. Given k p-values coming from k experiments, the combining function is based on the order of the p-values. For instance, a combination method might simply consist in taking the smallest of the p-values; another method, the second smallest, and so forth. They are not considered any further here because the more common forms would not produce p-values with the validity property.

On the other hand, quantile methods can satisfy this requirement. The quantile methods transform the p-values by using a function often chosen as the inverse of a Cumulative Distribution Function (CDF), which may and indeed generally does differ from that of the null hypothesis. The transformed values (which may be considered quantiles) are then added together and the aggregated p-value is computed using the CDF of the sampling distribution of the sum of those ”quantiles”. The choice of CDF is in principle arbitrary, but computational considerations constrain it to those distributions for which the calculations can be expressed with closed formulas or can be computed taking advantage of widely available tables. Combinations methods following the quantile framework have the property that if the p-values are uniformly distributed and independent to start with, their combination is uniformly distributed. This is necessary if the validity property of the CP is to be preserved. Here we consider two methods: Fisher’s method (also known as chi-square method) and Stouffer’s method (also known as z-transform test).

4.1. Fisher’s method

Fisher’s method (Fisher, 1932, 1948), also known as chi-square method, is among the earliest p-value combination methods. It relies on the key observation that if $p_1, p_2, \ldots, p_k$ are each the realization of a uniformly distributed random variable,

$$h_i = -2 \log p_i \qquad \text{with} \quad i = 1, \ldots, k$$

is a random variable that follows a chi-squared distribution with 2 degrees of freedom.

The sum of k independent random variables each following a chi-squared distribution with 2 degrees of freedom is itself chi-squared distributed with $2k$ degrees of freedom.

$$h = -2 \sum_{i=1}^{k} \log p_i$$

is a random variable that follows a chi-squared distribution with $2k$.

The combined p-value is:

$$p = \mathbb{P}\left\{ y \leq -2 \sum_{i=1}^{k} \log p_i \right\}$$

where y is a random variable following a chi-square distribution with $2k$ d.f. The integral required for calculating the probability above has a very simple closed form:

$$t \sum_{i=0}^{k-1} \frac{(-\log t)^i}{i!}$$

where $t = (p_1 \times p_1 \times \cdots \times p_k)$.

4.2. Stouffer's method

Stouffer's method (Stouffer et al., 1949), also known as z-transform method, maps the uniformly distributed p-values onto random variables with a normal distribution. This is achieved by:

$$h_i = \Phi^{-1}(1 - p_i)$$

where Φ is the cumulative normal distribution. If the p_i are independent, then:

$$h = \frac{\sum_{i=1}^{k} h_i}{\sqrt{k}}$$

is also normally distributed. The combined p-value is:

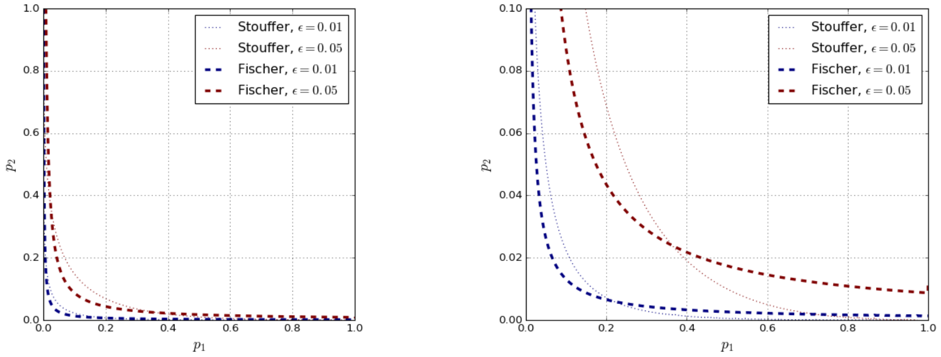
$$p = 1 - \Phi(h)$$

4.3. Comparison

As stated earlier, there is no method that is guaranteed to outperform all the others. A claim that is often cited is the Littell and Folks's proof (Littell and Folks, 1971, 1973) that ”Fisher's method is asymptotically optimal among essentially all methods of combining independent tests”, but the recurring advice from practitioner in the literature is to choose the method that best suits the characteristics of the evidence.

Figure 1 illustrates the rejection regions for the two methods for significance levels $\epsilon = 0.01$ and $\epsilon = 0.05$ when combining two p-values. Note that the contours for the two methods for the same significance levels intersect. This indicates that one method is not stricter than the other for all p-values.

Figure 1: The rejection regions for the Fisher method and the Stouffer method, for $\epsilon = 0.01$ and $\epsilon = 0.05$



5. Calibration to Bayes Factors

In the context of the discussions among probability theorists on the foundations of the notion of probability and more specifically on whether p-values can be really used as a measure of empirical evidence against a hypothesis (Berger and Sellke, 1987), a proposal has emerged to approach the combination of p-values by first transforming them into Bayes factors. For the present purposes, a Bayes factor is defined as:

$$B_\theta(x) = \frac{L_x(\theta)}{\int_\Theta L_x(\theta) dQ(\theta)}$$

where $L_x(\theta)$ is the likelihood of x given θ and $Q(\theta)$ a prior distribution in θ. The smaller a Bayes factor $B_\theta(x)$ is, the less likely it is that the parameter will take value θ having observed data x.

A p-value can be transformed into a Bayes factor by way of a *calibrator*. The reader is referred to (Vovk, 1993) and (Shafer et al., 2011) for the mathematical details. For the purposes of this paper, it will suffice to say that a non-decreasing and continuous function $f : (0, 1) \to (0, +\infty)$ is a calibrator if and only if

$$\int_0^1 \{1/f(p)\} dp \leq 1.$$

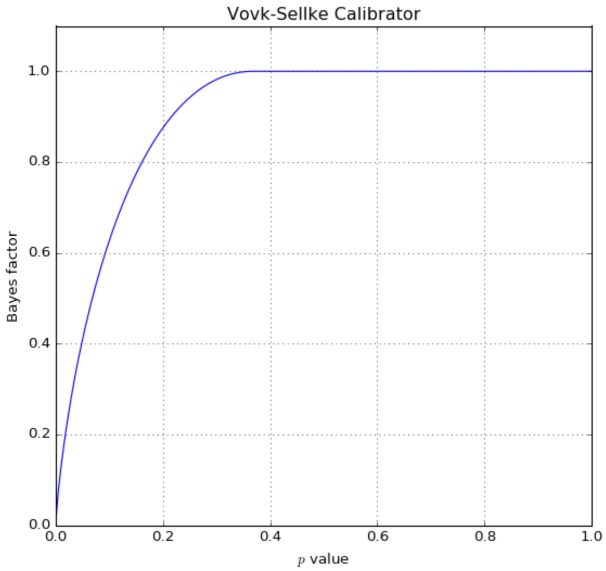
For instance, a family of calibrators is given by $f(p) := p^{1-\alpha}/\alpha$ for $\alpha \in (0, 1)$.

The calibrator that will be used in the empirical application is based on the Vovk-Sellke bound and has following form:

$$f(p) := \begin{cases} -ep\log(p) & p < 1/e \\ 1 & p \geq 1/e \end{cases}$$

The advantages accruing from this calibrator are discussed in (Bayarri et al., 2016).

Figure 2: The Vovk-Sellke calibrator



Having obtained Bayes factors, it is now possible to compute a combined p-value as:

$$p = \mathbb{P}\left\{ \prod_{i=1}^{k} f(q_i) \leq \prod_{i=1}^{k} f(p_i) \right\}$$

where the q_i are random variables uniformly distributed in $(0, 1)$ and p_i are the p-values to combine.

6. Empirical results

To assess the relative merits of the different approaches, the methods were applied on 3 sets of p-values, each obtained via Mondrian Inductive Conformal Predictors, using three different underlying ML algorithms, namely Neural Networks, SVM, and Random Forests. The initial data set was obtained from PubChem (Wang et al., 2017), a public repository of data on chemical compounds and biological assays. The data set (designated as AID827) was suggested by industry experts because its characteristics are representative of a large class of prediction problems in chemoinformatics. The data set is the product of a High Throughput Screening assay aimed at identifying chemical compounds that kill cells from a particular tumoral cell line[2]. The classification into Active vs. Inactive was carried out by applying a threshold on the estimated percentage of cells still alive after exposure to the chemical. The threshold was chosen by the suppliers of the data set, who also provided the associated classification.

[2] The complete designation of the data set is “High Throughput Screen to Identify Compounds that Suppress the Growth of Cells with a Deletion of the PTEN Tumor Suppressor”.

For the purposes of applying machine learning techniques, from each compound, a description of relevant features of its molecular structure was obtained via *signature descriptors* (Faulon et al., 2003). Each feature corresponds to the number of occurrences of a specific labelled subgraph in the labelled molecular graph of a compound. So, for each compound all the possible labelled subgraphs up to a chosen depth (max number of edges along a path from the root to a leaf) were enumerated and their occurrences counted. The key statistics of the resulting data set are summarized in Table 1 which shows the high imbalance, high sparsity, and high dimensionality common to many chemoinformatics prediction problems.

Table 1: Key statistics of the data set. The lower part refers to the data sets used in each of the 20 runs.

Total number of examples	=	138,437
Number of original features	=	170,334
Number of non-zero entries	=	7,868,562
Density of the data set	=	0.00034
Active compounds	=	1,659 (1.2%)
Number of selected features	=	6,262
Test objects	=	10,000
Calibration set size	=	10,000
Parameter optimization set size	=	10,000
Proper training set size	=	108,437

The dimensionality of data set was substantially reduced to keep the computational requirements manageable, especially for Neural Networks. The feature selection was performed very simply by filtering out all the features for which the variance (across all examples) was less than 0.001.

For the outcome to have some element of statistical significance, it was planned to repeat the evaluation 20 times. Consequently, for each of the 20 runs, the entire data set was randomly split into test set, calibration set, parameter optimization set, and proper training set. The split was stratified to ensure that the Active and the Inactive classes were represented in same proportions as in the original data set.

All the computations were run on the IT4I Salomon cluster and the Anselm cluster, both located in Ostrava, in the Czech Republic. The Salomon cluster is based on the SGI ICE X system and comprises 1008 computational nodes (plus a number of login nodes), each with 24 cores (2 12-core Intel Xeon E5-2680v3 2.5GHz processors) and 128GB RAM, connected via high-speed 7D Enhanced hypercube InfiniBand FDR and Ethernet networks. The Anselm cluster has 229 nodes, with a mixture of twin Intel 8-core 2.3GHz Sandy Bridge E5-2470 and twin Intel 8-core 2.4GHz Sandy Bridge E5-2665. 23 of the nodes have also one NVIDIA Tesla Kepler K20 GPU. Training and testing for each run was carried out on a single node, but runs were distributed across multiple nodes using the `dask/distributed` framework (Rocklin, 2016).

6.1. Algorithms

The algorithms were chosen with the aim of having inherently different approaches. Intuitively, ensembling in general and p-value combination in particular have a better chance of being beneficial if the component predictors complement each other in terms of predictive weaknesses and strengths.

6.1.1. Neural Networks

The architectural parameters of the Neural Network used in this experiment are captured in Table 2. The architecture is Feed-forward, the optimizer was Stochastic Gradient Descent, with a mini-batch size of 384. Dropout was applied with a rate of 0.80 on layer 2 (to prevent feature co-adaptation). The Tensorflow framework (Abadi et al., 2015) was used to implement the network and the model was trained on one node equipped with an NVidia K40 GPU. There was admittedly limited effort in optimizing the parameters and the topology. The convergence of the network was observed via Tensorboard, evaluating periodically the loss function on the parameter optimization set during training.

Table 2: Characteristics of Neural Network

	# nodes	**Activation function**	**Topology**
Input	6,262	—	—
Layer 1	2,048	ReLU	Fully connected
Layer 2	1,024	Tanh	Fully connected
Output	1	Sigmoid	Fully connected

One unusual aspect of the Neural Network in this exercise is the loss function used during training. It seemed intuitive that, to cater for the high imbalance, an asymmetric log-loss should be used. However the simple approach of assigning different weights to the two terms of the log-loss as in $L(p, y) = -w_0(1 - y)\log(1 - p) - w_1 y \log p$ leads to a loss function that is no longer proper. A proper loss function is such that $\mathbb{E}_{y \sim B_q} L(p, y)$, where B_q is the Bernoulli distribution with parameter q, attains its minimum at $p = q$. In other words, if y has a probability q of being 1, then the expectation of proper loss function as a function of p (fixed) is minimized for $p = q$. Informally, it is has been claimed that proper loss functions ”keep forecasters honest”. The proper form of an asymmetric log-loss with weights a and b was suggested in (Nouretdinov, 2016) and is:

$$L(p, y) = \begin{cases} -b\ \log(1-p) + (a-b)p & \text{if } y = 0 \\ -a\ \log(p) + (b-a)(1-p) & \text{if } y = 1 \end{cases}$$

Given the imbalanced class representation in the data set (the Active class is ≈ 1% of the total), the weights used during training were set to $a = 0.99$ and $b = 0.01$.

The NCM that was used for Conformal Prediction is

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

where $o()$ is the output of the neuron in the output layer.

6.1.2. Support Vector Machines

In this experiment, the SVM employed a kernel that is the composition of the Tanimoto kernel and the RBF kernel, as in previous experiments this seemed to be well suited to the specific task. A customized version of the very popular LIBSVM tool (Chang and Lin, 2011) was developed by one of the authors (Toccaceli, 2016) to allow for arbitrary kernels implemented for speed in C as external shared libraries. The parameters C, the weight for the active class, and γ (bandwidth of the RBF) were optimized once only (using the parameter optimization set), rather than for each of the runs.

The NCM is $-y_i f(x_i)$, where $f()$ is the decision function of the SVM and the labels are assumed to take values -1 (Inactive) and +1 (Active).

6.1.3. Random Forests

The implementation of Random Forests used in this investigation is the one in the `scikit-learn` Python package (Varoquaux et al., 2015). The RF consisted of 10000 fully grown trees. The trees were grown with the default setting of picking $\sqrt{p}$ random features (where p is the number of the features) at each stage. Also, the optimal split was chosen taking into account weights based on class representation in the training set.

The NCM chosen for RF was the fraction of trees that classified the test object as having the opposite label as the hypothetical one.

6.2. Classification Performance

The classification performance of the three algorithms is summarized in Figure 3. Given the high imbalance, accuracy is arguably not an appropriate metric. Instead, the performance was assessed in terms of Precision (fraction of Active test examples among the test example predicted as Active), Recall (fraction of all the Active test examples that were predicted as Active), and Area Under the ROC Curve (ROC AUC). In addition, the number of Uncertain predictions and the number of Empty predictions are also relevant metrics in this application of Conformal Predictors[3].

For this data set and for the parameter settings chosen in this study, NN and RF appear to share a common tendency to be more precise at the expense of recall, compared with SVM. All three algorithms achieve similar ROC AUC.

Mondrian Inductive Conformal Predictors were then applied, using the NCMs defined in the previous subsection. The resulting confusion matrices for the set predictor over the 10,000 test objects for each individual algorithm are reported in Table 3. The values in the table are the averages over the 20 runs. The rightmost column shows the count of the errors (actual label of the test object not in the prediction set); from this information, the validity property appears by and large verified (i.e. the number of errors is indeed roughly equal to significance level ϵ times the size of the test set, 10,000).

[3] Uncertain predictions occur when the Conformal Predictor outputs more than one label for the chosen level of significance. Empty predictions occur when the significance level is too high for the Conformal Predictor to output a label.

Table 3: Set Prediction Confusion Matrices for the Active class for each algorithm.

Neural Networks

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	38.05	97.95	0.55	705.55	0.00	9157.90	98.50
0.05	62.75	500.45	6.05	3122.70	0.00	6308.05	506.50
0.10	74.40	995.65	12.35	4750.35	0.00	4167.25	1008.00
0.15	81.90	1492.85	18.35	6052.05	0.00	2354.85	1511.20
0.20	87.60	1993.40	24.20	7016.55	3.85	874.40	2021.45
0.25	89.75	2156.60	25.90	7309.65	344.80	73.30	2527.30
0.50	58.40	387.85	13.00	4927.60	4613.15	0.00	5014.00
0.75	29.80	54.30	3.55	2468.15	7444.20	0.00	7502.05
0.80	24.10	35.25	2.50	1974.55	7963.60	0.00	8001.35
0.85	19.75	22.00	1.75	1473.05	8483.45	0.00	8507.20
0.90	12.60	10.85	1.05	992.35	8983.15	0.00	8995.05
0.95	6.95	4.90	0.30	506.35	9481.50	0.00	9486.70
0.99	1.95	0.85	0.05	106.90	9890.25	0.00	9891.15

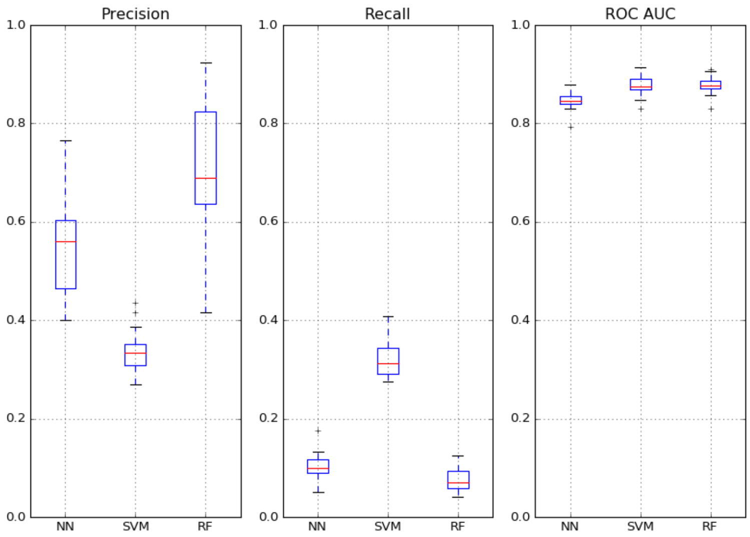
SVM (Tanimoto+RBF)

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	42.05	93.55	1.05	909.75	0.00	8953.60	94.60
0.05	69.85	490.75	6.75	3955.80	0.00	5476.85	497.50
0.10	83.65	991.10	12.55	5648.35	0.00	3264.35	1003.65
0.15	90.10	1478.30	18.65	6885.50	0.00	1527.45	1496.95
0.20	94.35	1866.00	22.70	7581.65	111.25	324.05	1999.95
0.25	91.10	1552.30	20.80	7401.90	933.90	0.00	2507.00
0.50	57.90	236.10	9.00	4937.05	4759.95	0.00	5005.05
0.75	28.35	37.15	2.85	2472.35	7459.30	0.00	7499.30
0.80	22.65	24.50	2.00	1978.65	7972.20	0.00	7998.70
0.85	17.05	16.00	1.30	1481.10	8484.55	0.00	8501.85
0.90	12.90	10.55	0.90	988.15	8987.50	0.00	8998.95
0.95	7.40	5.30	0.40	490.70	9496.20	0.00	9501.90
0.99	1.95	2.05	0.00	97.75	9898.25	0.00	9900.30

Random Forests

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	45.30	93.25	0.55	651.35	0.00	9209.55	93.80
0.05	68.75	490.20	6.50	4255.60	0.00	5178.95	496.70
0.10	81.80	989.35	13.15	6042.05	0.00	2873.65	1002.50
0.15	89.60	1483.75	18.90	7141.05	0.00	1266.70	1502.65
0.20	93.35	1794.65	23.75	7726.85	177.85	183.55	1996.25
0.25	90.00	1534.40	20.60	7407.30	947.70	0.00	2502.70
0.50	57.75	236.20	8.80	4962.50	4734.75	0.00	4979.75
0.75	29.60	27.90	2.20	2564.75	7375.55	0.00	7405.65
0.80	23.80	18.55	1.40	2072.05	7884.20	0.00	7904.15
0.85	18.15	12.45	0.90	1550.20	8418.30	0.00	8431.65
0.90	12.75	6.70	0.60	1133.30	8846.65	0.00	8853.95
0.95	6.85	2.70	0.35	649.15	9340.95	0.00	9344.00
0.99	1.85	0.45	0.15	325.10	9672.45	0.00	9673.05

Figure 3: Performance of the Neural Networks, SVM, and RF.



6.3. Performance of Fisher and Stouffer methods

The confusion matrices for Fisher and Stouffer methods are reported in Table 4 and Table 5, respectively. Each row contains the confusion matrix entries for one significance level value. The table should make it possible to choose the significance value that results in the Precision and Recall that best suit a specific application. Both Fisher and Stouffer methods result in better efficiency, as the number of uncertain predictions is reduced compared to any of the single-algorithm results. Within the same method, the efficiency appears to improve when combining 3 p-values compared to combining 2 p-values. However, validity is adversely affected for low values of the significance level (i.e. more errors than expected are made). The point is illustrated in more detail in Figure 4, which shows that the deviation from ideal validity is symmetrical for Stouffer's method, whereas it is asymmetrical for Fisher's method, with a smaller deviation for low ϵ and a more pronounced deviation (fewer errors than expected) elsewhere.

It should be noted that both Fisher's and Stouffer's method depend on the assumption of independence and of uniform distribution. Some researchers have proposed methods (Brown, 1975; Alves G., 2014; Poole et al., 2016) for mitigating the consequences of correlation, but experimentation with these methods has been left for further study.

In some applications it is advantageous to rank test objects according to how supportive the evidence is of them being of one class rather than the other. In the example used here, one may want to rank compounds by how strongly the evidence support their being Active for the biological target in hand. Note that there are two ways to do this: ranking the compounds by highest p_{active} or ranking them by lowest $p_{inactive}$. The latter is arguably more in line with the tenets of Statistical Hypothesis Testing: the compounds that rank at the top are those for which the hypothesis of them being Inactive can be rejected with stronger

Table 4: Set Prediction Confusion Matrices for the Active class after combining p-values with the Fisher method

Neural Networks + SVM

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	61.25	296.55	2.85	2450.90	0.00	7188.45	299.40
0.05	78.35	818.90	11.05	5131.45	0.00	3960.25	829.95
0.10	87.20	1299.00	16.15	6511.90	2.00	2083.75	1317.15
0.15	91.40	1699.75	21.45	7339.75	38.50	809.15	1759.70
0.20	92.15	1789.95	23.70	7641.60	350.55	102.05	2164.20
0.25	88.15	1464.30	21.30	7351.10	1073.50	1.65	2559.10
0.50	64.70	398.00	11.05	5480.90	4045.35	0.00	4454.40
0.75	42.90	93.90	5.05	3443.25	6414.90	0.00	6513.85
0.80	37.55	68.15	3.75	2968.95	6921.60	0.00	6993.50
0.85	31.70	45.55	2.65	2463.20	7456.90	0.00	7505.10
0.90	25.60	27.05	1.55	1889.15	8056.65	0.00	8085.25
0.95	17.30	13.90	0.80	1204.10	8763.90	0.00	8778.60
0.99	7.00	4.15	0.25	423.70	9564.90	0.00	9569.30

Neural Networks + SVM + RF

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	71.85	472.85	6.15	4139.75	0.00	5309.40	479.00
0.05	85.60	1027.95	14.95	6464.80	0.35	2406.35	1043.25
0.10	91.40	1475.05	20.40	7478.55	18.20	916.40	1513.65
0.15	93.45	1634.70	22.30	7820.95	256.55	172.05	1913.55
0.20	90.80	1405.05	21.45	7624.35	855.05	3.30	2281.55
0.25	86.70	1120.55	19.75	7298.70	1474.30	0.00	2614.60
0.50	67.35	390.00	11.00	5708.20	3823.45	0.00	4224.45
0.75	48.65	116.55	5.05	3940.45	5889.30	0.00	6010.90
0.80	44.75	87.40	4.20	3510.55	6353.10	0.00	6444.70
0.85	39.75	63.15	3.00	3021.20	6872.90	0.00	6939.05
0.90	33.65	41.15	1.85	2458.95	7464.40	0.00	7507.40
0.95	25.80	22.25	1.25	1729.00	8221.70	0.00	8245.20
0.99	13.80	7.70	0.25	762.90	9215.35	0.00	9223.30

evidence. This study examined the implications of p-value combination on the test object ranking. The results are reported in Table 6 and Table 7 for the $p_{inactive}$-based and p_{active}-based ranking, respectively. The tables show how many actually Active test compounds were listed among the 25 top ranked compounds. The bottom row shows that the p-value combination of NN and SVM results in a higher average count of Active compounds, for Fisher's as well as for Stouffer's methods. The 3-way combination of NN, SVM, and RF on the other hand improves on the performance of RF (and the other algorithms) only in the case of Fisher's method and when ranking by highest p_{active}. The detail of the tables allows to see also that combining is not always advantageous, even when on average it appears to be.

The statistical significance of the observed difference in the counts of Active compounds among the top 25 can be assessed with a paired observation test. The Wilcoxon signed-

Table 5: Set Prediction Confusion Matrices for the Active class after combining p-values with the Stouffer method

Neural Networks + SVM

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	65.60	400.95	5.00	3182.15	0.00	6346.30	405.95
0.05	82.80	1038.70	13.55	5891.25	0.00	2973.70	1052.25
0.10	90.55	1614.85	21.00	7234.55	1.95	1037.10	1637.80
0.15	92.25	1796.45	24.05	7702.35	312.45	72.45	2132.95
0.20	87.15	1369.90	20.95	7328.25	1193.75	0.00	2584.60
0.25	81.70	1016.60	18.95	6904.30	1978.45	0.00	3014.00
0.50	57.95	244.35	9.50	4986.00	4702.20	0.00	4956.05
0.75	35.90	63.60	3.90	3022.40	6874.20	0.00	6941.70
0.80	31.75	45.40	2.80	2585.75	7334.30	0.00	7382.50
0.85	27.20	30.30	1.80	2125.00	7815.70	0.00	7847.80
0.90	21.20	19.05	1.40	1619.05	8339.30	0.00	8359.75
0.95	14.30	10.40	0.90	1019.90	8954.50	0.00	8965.80
0.99	6.25	3.75	0.25	352.70	9637.05	0.00	9641.05

Neural Networks + SVM + RF

Significance level	Active pred Active	Inactive pred Active	Active pred Inactive	Inactive pred Inactive	Empty preds	Uncertain preds	Errors
0.01	77.85	677.35	9.70	5234.20	0.00	4000.90	687.05
0.05	90.45	1388.30	20.05	7384.70	0.10	1116.40	1408.45
0.10	93.15	1587.45	22.90	7887.75	376.40	32.35	1986.75
0.15	87.45	1157.35	20.40	7464.40	1270.40	0.00	2448.15
0.20	81.55	870.55	18.40	7071.85	1957.65	0.00	2846.60
0.25	77.15	665.05	16.00	6703.15	2538.65	0.00	3219.70
0.50	57.30	208.75	8.30	5064.40	4661.25	0.00	4878.30
0.75	40.30	65.80	3.55	3377.60	6512.75	0.00	6582.10
0.80	36.15	49.75	2.80	2983.05	6928.25	0.00	6980.80
0.85	31.40	35.60	2.00	2555.40	7375.60	0.00	7413.20
0.90	26.80	24.35	1.45	2080.40	7867.00	0.00	7892.80
0.95	19.90	14.55	1.00	1478.55	8486.00	0.00	8501.55
0.99	10.55	5.30	0.25	746.95	9236.95	0.00	9242.50

rank test (Wilcoxon, 1945; Hollander and Wolfe, 1999) is possibly a reasonable choice. The null hypothesis of the Wilcoxon signed-rank test is that the distribution of the differences between elements of pairs is symmetrical around 0. However, in its basic form, the test does not apply to variables with discrete values such as counts but only to variables with continuous values, the reason being that the test was not designed to deal (a) with no differences in a pair and (b) with ties among the differences (occurrences of pairs with the same difference in absolute value). Variants have been proposed (by Wilcoxon himself, who suggested to disregard the observation pairs with no difference, and by Pratt (Pratt, 1959), who suggested a way to account for those) but the distribution of the statistic would change. Simulations performed by one of the authors to study the effect of quantization (Toccaceli, 2017) appear to suggest that such variants are slightly conservative, in the sense that a value of the statistic that the Wilcoxon distribution would associate with $p = 1\%$ corresponds

Figure 4: Validity plot. This illustrates the deviation from validity introduced by Fisher and Stouffer Methods

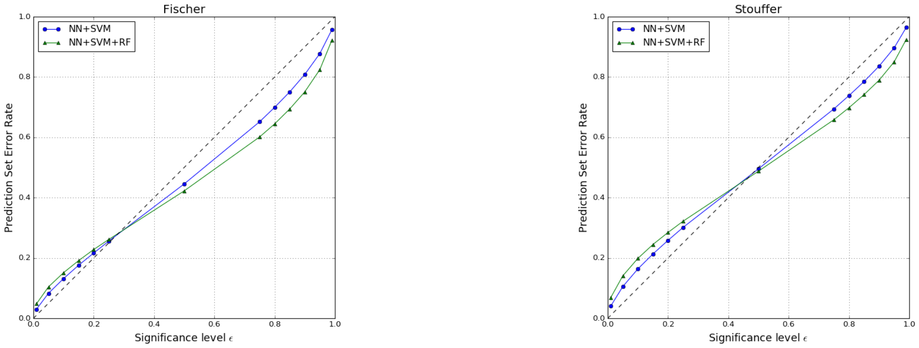


Table 6: Number of Active compounds among the top 25 test objects ranked by lowest $p_{inactive}$

				Fisher		Stouffer	
data set id	NN	SVM	RF	NN+SVM	NN+SVM+RF	NN+SVM	NN+SVM+RF
000	10	14	15	13	17	13	18
001	15	16	18	16	18	16	17
002	13	16	16	18	17	18	17
003	13	15	17	16	17	16	17
004	15	11	15	14	15	14	15
005	13	13	16	14	16	15	16
006	15	16	18	16	17	16	17
007	12	14	14	13	15	13	15
008	13	14	15	15	16	15	15
009	10	10	13	12	13	12	14
010	16	13	15	13	15	13	15
011	12	10	16	13	14	13	14
012	13	14	16	16	16	16	17
013	18	19	19	18	20	18	20
014	13	10	14	13	14	13	14
015	12	13	15	14	15	13	16
016	16	13	20	16	16	16	16
017	11	15	15	12	14	12	14
018	13	15	16	14	15	14	15
019	13	14	14	13	14	13	14
Average	13.30	13.75	15.85	14.45	15.70	14.45	15.80

Table 7: Number of Active compounds among the top 25 test objects ranked by highest p_{active}

				Fisher		Stouffer	
data set id	NN	SVM	RF	NN+SVM	NN+SVM+RF	NN+SVM	NN+SVM+RF
000	10	14	15	14	18	13	15
001	15	15	18	16	17	17	18
002	12	16	17	16	18	17	18
003	14	16	17	15	18	15	18
004	16	12	14	14	15	14	14
005	14	13	16	14	16	14	14
006	15	17	18	15	17	15	16
007	13	13	14	14	16	13	15
008	13	14	15	15	15	15	15
009	11	11	13	12	13	12	14
010	15	15	15	14	15	13	15
011	11	10	16	12	14	13	13
012	14	14	16	17	17	17	16
013	18	19	20	21	21	21	20
014	13	9	14	13	15	13	14
015	13	13	15	13	15	13	14
016	16	14	20	16	17	16	16
017	11	15	15	12	14	11	13
018	14	15	17	14	15	14	16
019	14	14	13	14	15	14	14
Average	13.60	13.95	15.90	14.55	16.05	14.50	15.40

in fact to a lower p for the variants and is therefore stronger evidence against the null hypothesis.

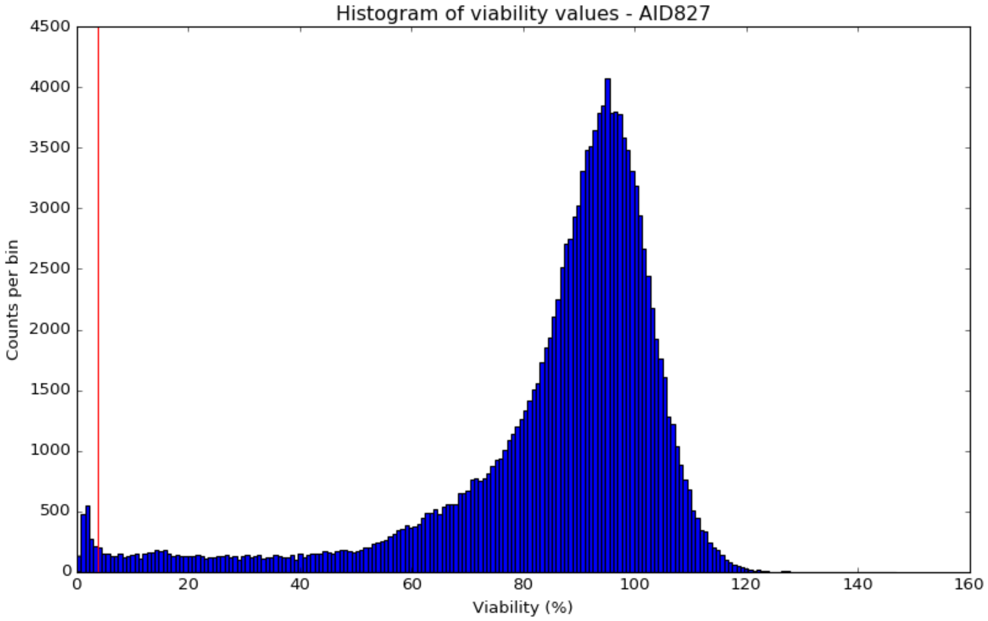
To get an indication of the statistical significance, it makes sense to limit the comparison to the best combination with its best component. The statistical significance between SVM and Fisher NN+SVM is 0.078 with plain Wilcoxon and 0.076 with the Pratt variant (computed with the `scipy` Python package (Jones et al., 2001)). So, while at first glance there appear to be an advantage, a result such as this one or a more convincing one could occur by chance (under the null hypothesis that the distribution of the difference is symmetrical) 1 out of 13 times. The difference in average between RF and Fisher NN+SVM+RF is very small and common sense alone is enough to surmise that the evidence does not contradict the hypothesis that the two are in fact the same. Just for completeness, the statistical significance in this case between is 0.75 with plain Wilcoxon and 0.252 with Pratt variant.

As a final observation, it may be worthwhile to take a look at the original data from which the data set used in this study was extracted. As explained in section 6, the label Active/Inactive was derived by thresholding a continuous value, referred to as *viability*, expressing the percentage of cells still alive after exposure to the compound. In particular, any compound for which the viability was less than or equal to 3.81% was deemed Active, otherwise Inactive[4]. A histogram of the Viability for data set AID827 is provided in Figure 5. Table 8 shows the top 25 compounds identified by highest p_{active} and by smallest $p_{inactive}$. Inspecting the viability, one realizes that several of the compounds classed as Inactive are

[4] In this specific assay, Activity denotes that the compound kills cells belonging to a specific tumoral cell line.

actually borderline cases. This occurrence is rather intriguing: while it is true that there are outright errors in the top 25, it is also true that the borderline cases are over-represented, suggesting that the classifiers did generalize on the data set and that the performance might in fact be better than the metrics on Active/Inactive classification indicate.

Figure 5: Histogram of Viability in AID827. The vertical line shows the value of the threshold.



6.4. Vovk-Sellke calibration

As illustrated in Figure 2, the calibrator assigns the same Bayes factor of 1 to p-values greater than $1/e$. The rationale is that in Statistical Hypothesis Testing one can assume that p-values above a certain value cease to be informative. The emphasis is on low values because these are what constitute strong evidence on which to reject the null hypothesis.

When applied to combining Conformal Predictors, the V-S calibrator inevitably affects validity for high values of p, as the confusion matrices in Table 9 and the chart in Figure 6 attest. The combined CP appears to predict with substantially fewer errors than the significance level would allow. Also, for lower values of p-values, on the other hand, the deviation from validity is limited and improves on either Fisher's or Stouffer's methods.

As to the performance on ranking, because of what was observed at the start of this subsection, the ranking by highest of test objects by largest p_{active} becomes meaningless and is reported here only for completeness. It is in the ranking of compounds by lowest $p_{inactive}$ that V-S calibration finds its appropriate application. Its averages of 14.50 for NN+SVM and 15.55 for NN+SVM+RF are in line with those of Fisher's and Stouffer's methods up to statistical fluctuations.

Table 8: Example of the top 25 compounds (from run 000, Stouffer NN+SVM+RF). The table on the left is order by lowest $p_{inactive}$, the one the right by highest p_{active}.

Rank	Compound tag	Viability	$p_{inactive}$
1	79813	1.76	3.483e-10
2	129543	4.57	9.419e-10
3	115173	1.48	1.593e-09
4	108813	15.69	2.372e-09
5	100523	0.85	4.316e-09
6	116614	39.05	2.161e-08
7	94529	3.57	2.312e-08
8	104764	1.47	3.455e-08
9	62991	25.27	4.058e-08
10	64246	4.44	4.743e-08
11	84878	1.77	4.755e-08
12	127825	1.67	5.238e-08
13	52454	2.95	5.885e-08
14	74599	3.84	6.941e-08
15	75236	74.03	9.263e-08
16	91399	2.05	1.138e-07
17	121411	1.69	1.929e-07
18	6106	2.27	2.118e-07
19	104197	1.78	2.127e-07
20	12551	1.08	2.363e-07
21	85895	2.03	2.412e-07
22	128112	1.96	2.579e-07
23	96373	1.16	2.599e-07
24	74016	2.37	2.820e-07
25	130880	3.36	3.077e-07

Rank	Compound tag	Viability	p_{active}
1	115173	1.48	1.000
2	116614	39.05	1.000
3	129543	4.57	1.000
4	79813	1.76	1.000
5	100523	0.85	0.998
6	108813	15.69	0.998
7	94529	3.57	0.997
8	62991	25.27	0.994
9	64246	4.44	0.992
10	84878	1.77	0.990
11	104764	1.47	0.988
12	127825	1.67	0.985
13	52454	2.95	0.984
14	74599	3.84	0.982
15	75236	74.03	0.978
16	115494	83.84	0.977
17	121411	1.69	0.977
18	91399	2.05	0.977
19	119648	80.08	0.973
20	128112	1.96	0.964
21	85895	2.03	0.961
22	129514	50.91	0.960
23	130880	3.36	0.958
24	6106	2.27	0.958
25	104197	1.78	0.957

Figure 6: Validity plot for the combination via Vovk-Sellke calibration

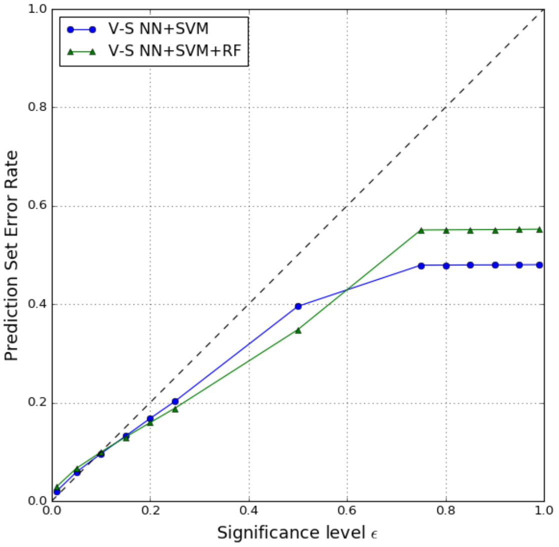


Table 9: Set Prediction Confusion Matrices for the Active class after combining p-values with the V-S Calibration method

Neural Networks + SVM

Significance	Active	Inactive	Active	Inactive	Empty	Uncertain	Errors
level	pred	pred	pred	pred	preds	preds	
	Active	Active	Inactive	Inactive			
0.01	55.25	198.55	1.45	1630.90	0.00	8113.85	200.00
0.05	72.25	578.45	7.50	4103.15	0.35	5238.30	586.30
0.10	81.10	944.55	12.60	5490.85	4.70	3466.20	961.85
0.15	87.00	1275.60	15.20	6343.05	29.50	2249.65	1320.30
0.20	89.95	1548.80	18.55	6942.90	111.00	1288.80	1678.35
0.25	90.95	1719.90	21.15	7291.55	282.25	594.20	2023.30
0.50	75.30	754.10	13.45	5968.00	3189.15	0.00	3956.70
0.75	65.25	418.10	10.15	5141.40	4365.10	0.00	4793.35
0.80	65.25	418.10	10.15	5140.10	4366.40	0.00	4794.65
0.85	65.25	418.10	10.15	5137.95	4368.55	0.00	4796.80
0.90	65.25	418.10	10.15	5137.00	4369.50	0.00	4797.75
0.95	65.25	418.10	10.15	5134.25	4372.25	0.00	4800.50
0.99	65.25	418.10	10.15	5131.30	4375.20	0.00	4803.45

Neural Networks + SVM + RF

Significance	Active	Inactive	Active	Inactive	Empty	Uncertain	Errors
level	pred	pred	pred	pred	preds	preds	
	Active	Active	Inactive	Inactive			
0.01	63.65	288.60	2.65	2681.95	0.00	6963.15	291.25
0.05	77.25	654.40	8.60	5044.45	0.55	4214.75	663.55
0.10	84.15	972.25	13.40	6188.40	9.35	2732.45	995.00
0.15	88.05	1237.50	16.80	6854.95	44.35	1758.35	1298.65
0.20	90.85	1441.85	18.95	7270.45	132.55	1045.35	1593.35
0.25	92.30	1560.35	20.50	7482.20	303.50	541.15	1884.35
0.50	81.50	922.90	15.40	6436.45	2543.75	0.00	3482.05
0.75	60.55	291.50	7.05	4431.20	5209.70	0.00	5508.25
0.80	60.55	291.50	7.05	4428.05	5212.85	0.00	5511.40
0.85	60.55	291.50	7.05	4424.85	5216.05	0.00	5514.60
0.90	60.55	291.50	7.05	4422.65	5218.25	0.00	5516.80
0.95	60.55	291.50	7.05	4419.15	5221.75	0.00	5520.30
0.99	60.55	291.50	7.05	4415.30	5225.60	0.00	5524.15

7. Conclusions and future work

This study discussed different methods for combining p-values produced by Conformal Predictors. The methods chosen here arise from considerations belonging to statistical hypothesis testing rather than statistical learning proper and their computational cost is next to negligible (in the order of fraction of a second for each of the data sets used here). The study demonstrated on a real-world example that, despite their simplicity, these techniques can be of benefit, in particular with the Fisher method exhibiting a synergistic effect on the accuracy of ranking as in the case of the combination of NN and SVM. In the tests, while there was no evidence that the benefits extend to multiple combinations, there was also no evidence of negative effects. The deviation from validity of the set predictor was also limited and combination appeared to improve efficiency.

Table 10: Number of Active compounds among the top 25 test objects after combining p-value via V-S calibration

By lowest $p_{inactive}$

run	NN	SVM	RF	NN	NN
				SVM	SVM
					RF
000	10	14	15	13	16
001	15	16	18	16	18
002	13	16	16	18	17
003	13	15	17	16	17
004	15	11	15	14	15
005	13	13	16	14	16
006	15	16	18	16	17
007	12	14	14	13	15
008	13	14	15	15	15
009	10	10	13	12	13
010	16	13	15	13	15
011	12	10	16	13	14
012	13	14	16	16	16
013	18	19	19	19	20
014	13	10	14	13	14
015	12	13	15	14	15
016	16	13	20	16	16
017	11	15	15	12	13
018	13	15	16	14	15
019	13	14	14	13	14
Average	13.30	13.75	15.85	14.50	15.55

By highest p_{active}

run	NN	SVM	RF	NN	NN
				SVM	SVM
					RF
000	10	14	15	1	8
001	15	15	18	4	4
002	12	16	17	1	4
003	14	16	17	1	6
004	16	12	14	5	5
005	14	13	16	2	4
006	15	17	18	3	8
007	13	13	14	3	7
008	13	14	15	4	3
009	11	11	13	5	7
010	15	15	15	1	3
011	11	10	16	3	5
012	14	14	16	1	3
013	18	19	20	7	5
014	13	9	14	6	7
015	13	13	15	5	8
016	16	14	20	1	6
017	11	15	15	4	3
018	14	15	17	0	3
019	14	14	13	6	4
Average	13.60	13.95	15.90	3.15	5.15

One possible future line of enquiry might be about intelligent ways of mixing the p-values on the basis of the objects to which they refer. The methods discussed so far rely only on the bare p-values. They do not exploit any patterns in the different accuracy of the different underlying ML of Conformal Predictors. One Conformal Predictor might be more accurate than the others in one range of predicted values, but not in another. One CP could be systematically more accurate for some subsets of object, whereas another CP might be more accurate for a different subset. One way to try to exploit these different abilities might be by learning which objects tends to be better predicted by which CP. *Mixture of Experts* models (Jacobs et al., 1991) use a combination of specialized models and a gating network which weights, possibly in a non-linear way, the output of the specialized models. The gating network uses as inputs the objects, their labels and the predictions of the models. In such a framework, scalability could be achieved by partitioning the data set across multiple nodes and then aggregating the p-values. In the specific chemoinformatics problem used as an example here, it may even make sense to have component classifiers becoming specialized by assigning training examples from the same chemical cluster. This approach could also allow to frame the p-value combination as an optimization problem over an appropriate functional space (such as a RKHS), with constraints to enforce validity and with a loss function crafted to improve efficiency.

8. Acknowledgments

This project (ExCAPE) has received funding from the European Unions Horizon 2020 Research and Innovation programme under Grant Agreement no. 671555. We are grateful for the help in conducting experiments to the Ministry of Education, Youth and Sports (Czech Republic) that supports the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center LM2015070". This work was also supported by EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots") and by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK. We are indebted to Lars Carlsson of AstraZeneca for providing the data and for the useful discussions. We are also thankful to Zhiyuan Luo, Vladimir Vovk, and Ilia Nouretdinov for many valuable comments and suggestions.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

Yu Y.K. Alves G. Accuracy evaluation of the unified p-value from combining correlated p-values. *PloS one*, 9(3), 2014. doi: 10.1371/journal.pone.0091225.

M.J. Bayarri, Daniel J. Benjamin, James O. Berger, and Thomas M. Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90 – 103, 2016. ISSN 0022-2496. doi: http://dx.doi.org/10.1016/j.jmp.2015.12.007. URL http://www.sciencedirect.com/science/article/pii/S002224961600002X. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments.

James O. Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence (c/r: p123-133, 135-139, 1201-1201). *Journal of the American Statistical Association*, 82:112–122, 1987.

Morton B. Brown. A method for combining non-independent, one-sided tests of significance (corr: V32 p955). *Biometrics*, 31:987–992, 1975.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Ori Davidov. Combining p-values using order-based methods. *Computational Statistics & Data Analysis*, 55(7):2433–2444, 2011. doi: 10.1016/j.csda.2011.01.024. URL http://dx.doi.org/10.1016/j.csda.2011.01.024.

Jean-Loup Faulon, Donald P. Visco Jr., 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. doi: 10.1021/ci020345w. URL http://dx.doi.org/10.1021/ci020345w.

R. A. Fisher. Question 14: Combining independent tests of significance. *The American Statistician*, 2(5):30–30, 1948.

R.A. Fisher. *Statistical methods for research workers, 4th. ed.* Edinburgh Oliver & Boyd, 1932.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning: The second *Computer Journal* lecture. *Comput. J.*, 50(2):151–163, 2007. doi: 10.1093/comjnl/bxl065. URL http://dx.doi.org/10.1093/comjnl/bxl065.

Myles Hollander and Douglas A Wolfe. *Nonparametric statistical methods*. Wiley Series in Probability and Statistics. Wiley, New York, NY, 1999.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, March 1991. ISSN 0899-7667. doi: 10.1162/neco.1991.3.1.79. URL http://dx.doi.org/10.1162/neco.1991.3.1.79.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL http://www.scipy.org/. [Online; accessed 2017-04-09].

Ramon C. Littell and J. Leroy Folks. Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66:802–806, 1971.

Ramon C. Littell and J. Leroy Folks. Asymptotic optimality of Fisher's method of combining independent tests. II. *Journal of the American Statistical Association*, 68:193–194, 1973.

Thomas M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004.

Ilia Nouretdinov. personal communication, 2016.

William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard, and Theo A. Knijnenburg. Combining dependent P-values with an empirical adaptation of brown's method. *Bioinformatics*, 32(17):430–436, 2016. doi: 10.1093/bioinformatics/btw438. URL http://dx.doi.org/10.1093/bioinformatics/btw438.

John W. Pratt. Remarks on zeros and ties in the Wilcoxon signed rank procedure. *Journal of the American Statistical Association*, 54:655–667, 1959.

Matthew Rocklin. dask/distributed: Distributed computation in python, 2016. URL https://github.com/dask/distributed. [Online; accessed 2017-04-09].

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, June 2008. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1390681.1390693.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, bayes factors, and p-values. *Statistical Science*, 26(1):84–101, 2011.

E.A. Stouffer, S.A.and Suchman, L.C. DeVinney, S.A. Star, and R.M. Jr. Williams. *The American Soldier, Vol.1: Adjustment during Army Life.* Princeton University Press, Princeton, 1949.

Paolo Toccaceli. Experimental extension of libsvm, supporting arbitrary kernels via dynamically loaded library, 2016. URL https://github.com/ptocca/libsvm. [Online; accessed 2017-04-09].

Paolo Toccaceli. personal communication, 2017. URL http://clrc.rhul.ac.uk/people/ptocca/Reports/20170227-WilcoxonReport.pdf.

Paolo Toccaceli, Ilia Nouretdinov, and Alexander Gammerman. Conformal predictors for compound activity prediction. In Alexander Gammerman, Zhiyuan Luo, Jesús Vega, and Vladimir Vovk, editors, *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings*, pages 51–66. Springer International Publishing, Cham, 2016. ISBN 978-3-319-33395-3. doi: 10.1007/978-3-319-33395-3_4. URL http://dx.doi.org/10.1007/978-3-319-33395-3_4.

Gaël Varoquaux, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. Scikit-learn: Machine learning without learning the machinery. *GetMobile*, 19(1):29–33, 2015. doi: 10.1145/2786984.2786995. URL http://doi.acm.org/10.1145/2786984.2786995.

V. G. Vovk. A logic of probability, with application to the foundations of statistics (disc: p341-351). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 55: 317–341, 1993.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.

Yanli Wang, Stephen H. Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He, and Jian Zhang. Pubchem bioassay: 2017 update. *Nucleic Acids Research*, 45(Database-Issue):D955–D963, 2017. doi: 10.1093/nar/gkw1118. URL http://dx.doi.org/10.1093/nar/gkw1118.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.