Prediction with Confidence Based on a Random Forest Classifier

Dmitry Devetyarov and Ilia Nouretdinov

Computer Learning Research Centre, Royal Holloway, University of London,
Egham, Surrey, UK
{dmitry,ilia}@cs.rhul.ac.uk

Abstract. Conformal predictors represent a new flexible framework that outputs region predictions with a guaranteed error rate. Efficiency of such predictions depends on the nonconformity measure that underlies the predictor. In this work we designed new nonconformity measures based on a random forest classifier. Experiments demonstrate that proposed conformal predictors are more efficient than current benchmarks on noisy mass spectrometry data (and at least as efficient on other type of data) while maintaining the property of validity: they output fewer multiple predictions, and the ratio of mistakes does not exceed the preset level. When forced to produce singleton predictions, the designed conformal predictors are at least as accurate as the benchmarks and sometimes significantly outperform them.

Key words: Conformal predictor, confidence machine, region prediction, random forest

1 Introduction

The new framework of conformal prediction introduced in [1] allows us to output region predictions (a set of predicted labels) with the guaranteed error rate under a simple statistical assumption (this property is called *validity*), as opposed to point predictions when we always produce singleton predictions but the error rate is not guaranteed.

Having a guaranteed error rate, we may still obtain multiple region predictions, that is, predictions that comprise more than one label. Multiple predictions are not mistakes: they indicate that there was no sufficient information provided for predicting one label. The question is how big these region predictions are, and our aim is to decrease a number of multiple predictions. The ability of conformal predictors to produce predictions as certain as possible is called *efficiency*.

Most of known machine learning algorithms can be used as an *underlying algorithm* in a conformal predictor. Efficiency of conformal predictors is usually in line with the accuracy of the underlying algorithm and therefore varies across the range of underlying algorithms and also depends on the type of data analysed. For this reason, we are looking for new nonconformity measures that could result in efficient predictions.

So far various nonconformity measures have been designed. In this paper we propose new conformal predictors based on a random forest. We expect the predictors to inherit random forest advantages and to maintain the property of validity.

Although the main aim of this paper is to elaborate and analyse new methodology of providing region predictions, we can force conformal predictors to output one prediction instead of multiple predictions. Such approach allowed us to compare designed conformal predictors with machine learning methods (random forest, in particular). But one should bare in mind that in this case we do not have the advantages of conformal predictors as region predictors: there is no guaranteed validity.

2 Outline of Conformal Prediction

The framework of conformal prediction is described in detail in [1]. Conformal predictors are based on the only assumption about the data generating mechanism: all the examples have been generated independently by some probability distribution (the i.i.d. assumption).

Let us assume that we are given a training set of examples (x_1, y_1), $\ldots$, (x_{n-1}, y_{n-1}), where $x_i \in X$ is a vector of attributes and $y_i \in Y$ is a label out of a finite set of possible labels (classes), and our goal is to predict the classification y_n for remaining example x_n. A combination of an example and a label $z_i = (x_i, y_i) \in Z = X \times Y$ is called an object.

A nonconformity measure is a set of measurable mappings $\{A_n : n \in N\}$ of the type $A_n : Z^{(n-1)} \times Z \to (-\infty, +\infty]$, where $Z^{(n-1)}$ is the set of all bags (multisets) of elements of Z of size $n-1$. For each possible label y, we consider the hypothesis $y_n = y$ and the nonconformity measure assigns some values α_i (*nonconformity scores*) to every example in the sequence $\{z_i, i = 1, \ldots, n\}$ including a new example and evaluates 'nonconformity' $\alpha_i := A_n(\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int, z_i)$, $i = 1, \ldots, n$ between a set and its element ($\wr \ldots \int$ denotes a multiset).

For each hypothesis $y_n = y$, we compare α_n to the other α_is and calculate $p(y) = |\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}|/n$ — the p-value associated with the possible label y for x_n. Thus, we can compliment each label with a p-value that shows how well a new example with this label conforms with the rest of the sequence.

The *conformal predictor* determined by the nonconformity measure $A_n, n \in N$ and a significance level ϵ is then defined as a measurable function $\Gamma : Z^* \times X \times (0,1) \to 2^Y$ (2^Y is a set of all subsets of Y) such that the prediction set $\Gamma^{(\epsilon)}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ is defined as the set of all labels $y \in Y$ such that $p_n > \epsilon$. Thus, for any finite sequence of examples with labels, $(x_1, y_1, \ldots, x_{n-1}, y_{n-1})$, a new unlabelled object x_n and a significance level ϵ, the conformal predictor outputs a region prediction $\Gamma^{(\epsilon)}$ — a set of possible labels for a new object.

Conformal predictors defined above are *valid*: in the long run the frequency of errors made by a conformal predictor (that is, cases when prediction set Γ^ϵ

does not contain a true label) does not exceed ϵ subject to the i.i.d. assumption. Strictly speaking, for any exchangeable probability distribution P on Z^{∞} (Z^{∞} is the set of all infinite sequences of elements of Z) and any significance level ϵ,

$$\limsup_{n \to \infty} \frac{\sum_{i=1}^{n} err_n^{\epsilon}(\Gamma)}{n} \leq \epsilon \tag{1}$$

with probability one, where $err_n^{\epsilon}(\Gamma)$ is equal to 1 when the prediction set Γ^{ϵ} does not contain a real label y_n, and 0 otherwise. The property of validity is theoretically proven in *the on-line mode* and empirically confirmed in the *off-line mode* [1].

3 Random Forests

In this work we consider the type of random forests described in [4]. Theoretical results [4] demonstrate that random forests do not overfit when more trees are added. They also empirically proved to have the following advantages ([4], [5]): random forests produce high accuracy for many data sets; they can process data with a large number of features where each feature is weak, that is, carries a small amount of information; they are relatively robust to mixed variable types, missing data, outliers and noisy data; constructing random forests is relatively fast (faster than bagging and boosting).

In brief, a random forest is a classifier that consists of decision trees, each of which provides a vote for a certain class. Combining a large number of trees in a random forest can lead to more reliable predictions, while single decision tree may overfit the data.

3.1 Nonconformity Measures Based on Random Forests

In this paper we designed nonconformity measures based on random forests.

We will use the following notation: suppose we are given a bag $\wr(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\wr$, $(x_i, y_i) \in Z$, and we need to define a nonconformity measure $A(x, y) = A(\wr(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\wr; (x, y))$. Alternatively, we can define a conformity measure $B(x, y) = 1 - A(x, y)$ when it is more intuitive.

The nonconformity or conformity measures we propose are the following:

1. A random forest is constructed from a training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. The conformity score of a new example (x, y) is then equal to the percentage of correct predictions given for x by decision trees.

2. Conformity measure 1 is the most natural one, however, it is computationally inefficient: when considering example $N + 1$ we have to construct $(N + 1)L$ random forests, where L is the number of labels. We will therefore use another conformity measure, which will require only one random forest when making a prediction for a new object. The random forest is grown for the union of a bag $\wr(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\wr$ and a new example (x, y). Since for each decision tree, the training set is a bootstrap sample, a new example is not included

in this training set in about one third of decision trees. For each (x, y) we aggregate the votes for this example only of those decision trees where this example is out-of-bag (not in the training set for the tree). The conformity score is then equal to the proportion of correct votes for (x, y) among these trees.

3. This nonconformity measure was proposed by Huazhen Wang and Fan Yang in our personal communication. It is based on random forest proximities $P(i, j), i, j = 1, \dots, m + 1$, which provide a measure of how close to each other two objects are regardless of their labels and are calculated as a ratio of trees, running through which objects i and j land at the same terminal node. We construct a random forest for the union of a bag $\wr(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\wr$ and a new example (x, y) and form the corresponding $(m+1) \times (m+1)$ matrix of proximities for objects $x_1, x_2, \dots, x_m, x_{m+1} = x$. The nonconformity measure is the ratio of the average proximity of the example with examples of other classes to the average proximity of the example to examples of the same class. In both averages we consider only proximities of those k examples that have the greatest values of proximities among examples of the same class y and among all the other examples. Strictly speaking, $A(x, y) = A(x, y)^- / A(x, y)^+$, where

$$A(x,y)^+ = \sum_{s=1}^{k} P(i_s, m+1), A(x,y)^- = \sum_{s=1}^{k} P(j_s, m+1) \ , \tag{2}$$

i_s and j_s are the numbers of examples with s-st greatest value of proximity with example (x, y) among examples labelled with the same label y and among all the other examples, respectively.

4 Experiments

The designed nonconformity measures were implemented and applied to different data sets.

4.1 Data

In our experiments we used six proteomic data sets, two medical non-proteomic data sets and two non-medical data sets. Proteomic data sets comprise: ovarian cancer data from the *UKOPS* trial [6]; ovarian cancer (*OC*), breast cancer (*BC*) and heart disease (*HD*) samples collected in the *UKCTOCS* trial[1]; *Competition* data provided by the Leiden Clinical Mass Spectrometry Proteomic Diagnosis Competition and preprocessed as described in [7]; Ciphergen's *7 biomarkers* of UKOPS data [8]. The other medical data sets are shortened *Abdominal pain* data [9], which comprises 33 symptoms of acute abdominal pain, and *Microarray* data of lung cancer, colon cancer and breast cancer patients provided in the ICMLA 2009 Challenge [10]. *Sonar* and *Iris* non-medical data sets were taken from the University of California, Irvine (UCI) Machine Learning Repository.

[1] for more information see `www.ukctocs.org.uk`

4.2 Implemented Conformal Predictors

We implemented non-conformity measures 2 (referred to as *CP-RF*) and 3 (referred to as *CP-RF-kNN*, where k is the number of nearest neighbours). These conformal predictors were compared with benchmark predictors based on the kNN algorithm [3] (we will denote them *CP-kNN*, where k is the number of nearest neighbours) and SVM with different kernels [2] (denoted as *CP-SVM* (*kernel*, *parameter*)).

The experiments were carried out in two settings: off-line in leave-one-out (LOO) procedure, to show the usage of conformal predictors as conventional classifiers, and on-line, to demonstrate the advantages of region predictions.

We used the following parameters for random forest construction: the number of trees 1000, the number of features selected at each node to split equals a square root of the number of features. These values are recommended in [4], where it is theoretically proven that the results converge when we increase the number of trees in random forests and it is empirically shown that the results are insensitive to the number of features selected at each node.

4.3 Results

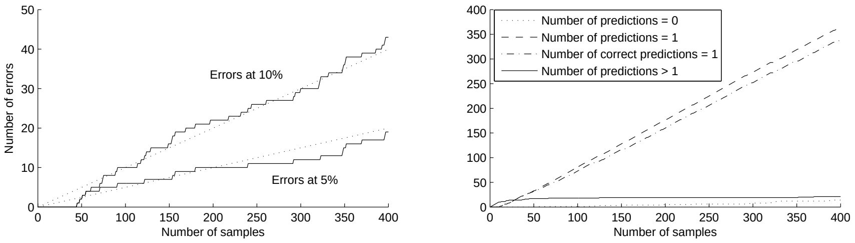
For each significance level $\epsilon > 0$, conformal predictors output a set of labels with p-values greater than ϵ. Thus, the predictor may output no label (we call this an *empty prediction*), one label (*certain prediction*) or more than one label (*multiple prediction*).

Firstly, the designed conformal predictors proved to be valid, that is, for a given significance level $\epsilon > 0$ the rate of erroneous predictions, that is, predictions not containing an actual label, is close to ϵ.

The example of the erroneous prediction dynamics is shown in Figure 1a. The figure demonstrates validity of the CP-RF-1NN applied to the Microarray data for significance levels $\epsilon = 5\%$ and 10% : solid lines, which represent the actual number of errors, are close to dotted lines, which demonstrate the expected number of errors for different significance levels.

Figure 1b demonstrates the dynamics of efficiency characteristics at significance level of 10% of the CP-RF-1NN applied to the Microarray data in the on-line mode. The characteristics shown are the number of multiple predictions, the number of certain predictions and the number of empty predictions. The figure demonstrates that while the number of analysed examples is low, they do not carry enough information to make certain predictions without losing validity. But starting from example 50, we have accumulated enough information so that multiple predictions cease to occur and most of prediction regions contain exactly one label, which is in most cases correct. The dynamics on the plot also conforms with the empirical fact established in [1] that when multiple predictions disappear, empty predictions start to occur.

As mentioned before, all implemented conformal predictors have a theoretically proven property of validity, and the general aim is to design a nonconformity measure that could improve efficiency, that is, make the algorithm output as few



(a) Validity (b) Efficiency

Fig. 1: Validity and efficiency of CP-RF-1NN applied to the Microarray data in the on-line mode.

multiple predictions and as many empty predictions as possible. Comparison of efficiency of CP-RF and CP-RF-kNN with the benchmarks demonstrated that CP-RF and CP-RF-kNN produce at least as few multiple predictions and as many correct certain and empty predictions as the benchmarks, and they perform much better in terms of efficiency on all mass spectrometry data sets. This allows us to speculate that conformal predictors based on random forests benefit from the advantages of the underlying algorithm and perform well on noisy data and data with a lot of weak inputs. Table 1 summarizes the multiple prediction rate for different conformal predictors.

Table 1: The rate of multiple predictions for significance level $\epsilon = 10\%$ in the LOO mode.

Data	CP-RF	CP-RF	CP-RF	CP-1NN	CP-5NN	CP-SVM	CP-SVM
		-1NN	-5NN			(rbf, 5)	(poly, 5)
UKOPS	46.1%	47.0%	45.8%	74.8%	72.0%	59.2%	69.8%
UKCTOCS OC	16.0%	16.0%	13.8%	44.6%	30.8%	38.5%	79.8%
UKCTOCS BC	77.8%	78.4%	77.8%	80.9%	80.9%	81.5%	82.7%
UKCTOCS HD	56.0%	58.1%	57.2%	64.7%	59.5%	66.1%	64.0%
Competition	11.1%	18.3%	17.0%	26.8%	19.6%	32.7%	30.7%
7 biomarkers	51.1%	55.4%	53.8%	67.0%	61.2%	97.9%	96.9%
Abdominal pain	0.3%	1.0%	0.0%	3.0%	0.0%	0.0%	1.0%
Microarray	0.0%	1.5%	0.3%	13.5%	3.5%	8.5%	40.4%
Sonar	14.9%	11.1%	13.0%	13.9%	16.4%	32.2%	30.8%
Iris	0.0%	0.0%	0.0%	0.0%	0.0%	86.7%	8.0%

Conformal predictors have been developed to provide region predictions with the preset error rate. However, in order to compare them with bare predictions output by conventional machine learning methods, we can ignore the nature of conformal predictors and force them to always make a certain prediction. After assigning a p-value for each label to every object, we can predict the label with the highest p-value. This is called *forced point prediction.* If several labels have the highest p-value (we call this situation a *tie*), we make a random prediction.

Experiments demonstrated that when forced to make point predictions, conformal predictors perform similarly to random forest algorithm (see Table 2). This can be explained by the fact that each random forest is a combination of a large number of trees constructed randomly and each sample is not included in the training set for about one third of all trees in a random forest.

This implies that we can add the framework of conformal prediction to the random forest algorithm without losing in accuracy, while benefiting from conformal predictions: we can produce valid region predictions and compliment each prediction with confidence.

The results of comparison of forced point prediction accuracy of different conformal predictors (Table 2) were in line with efficiency comparison: CP-RF and CP-RF-kNN significantly outperformed other predictors on certain mass spectrometry datasets and were at least as good as the benchmarks on all data sets.

Table 2: Accuracy of forced point predictions in the LOO mode.

Data	RF	CP-RF	CP-RF	CP-RF	CP-1NN	CP-5NN	CP-SVM	CP-SVM
			-1NN	-5NN			(rbf, 5)	(poly, 5)
UKOPS	72.6%	72.3%	71.5%	72.6%	55.1%	61.7%	66.7%	55.5%
UKCTOCS OC	84.9%	84.8%	83.8%	84.6%	72.3%	80.6%	78.9%	77.9%
UKCTOCS BC	66.0%	66.7%	59.0%	62.4%	50.3%	62.4%	56.2%	54.3%
UKCTOCS HD	71.8%	72.3%	69.2%	71.4%	63.2%	67.9%	62.4%	62.1%
Competition	83.7%	85.3%	83.7%	83.3%	82.0%	84.6%	86.3%	87.6%
7 biomarkers	74.6%	74.8%	72.2%	73.9%	64.5%	73.7%	60.9%	59.0%
Abdominal pain	91.7%	92.7%	91.8%	91.5%	88.0%	92.2%	91.7%	90.7%
Microarray	92.0%	91.3%	92.8%	91.4%	86.1%	89.4%	88.3%	89.5%
Sonar	85.1%	84.6%	88.7%	85.6%	86.3%	82.9%	84.6%	85.3%
Iris	95.3%	94.7%	95.0%	95.3%	93.3%	97.0%	89.3%	89.7%

5 Conclusion

In this paper we worked on further development of conformal predictors, which produce region predictions that make a preset number of errors in the long run. Designed nonconformity measures based on random forests proved to be valid and efficient. First, the ratio of mistakes does not exceed the preset level. Second, CP-RF and CP-RF-kNN produce more efficient predictions on all mass spectrometry data sets and are not beaten on the other data.

When forced to produce singleton predictions, conformal predictors based on random forest result in accuracy similar to random forest accuracy. The accuracy of forced point predictions output by CP-RF and CP-RF-kNN is at least as high as accuracy produced by known conformal predictors and sometimes is significantly higher. This implies that although conformal predictors are designed for producing valid region prediction, they can also be a useful tool when making singleton predictions.

Acknowledgments. We would like to thank our collaborators in MRC project ("Proteomic analysis of the human serum proteome") Ian Jacobs, Usha Menon, Rainer Cramer, John F. Timms, Ali Tiss, Jeremy Ford, Stephane Camuzeaux, Aleksandra Gentry-Maharaj for collecting the original UKCTOCS and UKOPS data and carrying out MS experiments.

This work is supported by EPSRC grant EP/F002998/1 (Practical competitive prediction); MRC grant G0301107 (Proteomic analysis of the human serum proteome); VLA of DEFRA grant on Development and Application of Machine Learning Algorithms for the Analysis of Complex Veterinary Data Sets; EU FP7 grant O-PTM-Biomarkers HEALTH-2007-2.4.1-2; ASPIDA grant "Development of new methods of conformal prediction with application to medical diagnosis" from the Cyprus Research Promotion Foundation; European Union EU FP7 programme "Living with uninvited guests: comparing plant and animal responses to endocytic invasions" (ERASysBio); and MRC Award G0802594.

References

1. Vovk, V., Gammerman, A., Shafer,G.: Algorithmic Learning in a Random World. Springer, New York (2005)
2. Gammerman, A., Vovk, V., Vapnik, V.: Learning by Transduction. In: 14th Conference on Uncertainty in Artificial Intelligence, pp. 148–155 (1998)
3. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive Confidence Machines for Pattern Recognition. Technical report 01-02, Royal Holloway, University of London (2001)
4. Breiman, L.: Random Forests. Mach Learn. 45, 5–32 (2001)
5. Breiman, L., Cutler, A.: Random Forests, `http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm#intro`
6. Timms, J.F., Cramer, R., Camuzeaux, S., Tiss, A., Smith, C., Burford, B., Nouretdinov, I., Devetyarov, D., Gentry-Maharaj, A., Ford, J., Luo, Z., Gammerman, A., Menon, U., Jacobs, I.: Peptides Generated Ex Vivo from Abundant Serum Proteins by Tumour-Specific Txopeptidases are Not Useful Biomarkers in Ovarian Cancer. Clin Chem. 56, 262–271 (2010)
7. Gammerman, A., Nouretdinov, I., Burford, B., Chervonenkis, A., Vovk, V., Luo, Z.: Clinical Mass Spectrometry Proteomic Diagnosis by Conformal Predictors. Stat Appl Genet Mo B. 7(2), Art. 13 (2008)
8. Nouretdinov, I., Burford, B., Luo, Z., Gammerman, A.: Data Analysis of 7 Biomarkers. Technical report, Royal Holloway, University of London (2008)
9. Gammerman, A., Thatcher, A.R.: Bayesian Diagnostic Probabilities without Assuming Independence of Symptoms. Method Inform Med. 30(1), 15–22 (1991)
10. Nouretdinov, I., Burford, B., Gammerman, A.: Application of Inductive Confidence Machine to ICMLA Competition Data. In: The Eighth International Conference on Machine Learning and Applications, pp. 435–438 (2009)