Reverse conformal approach for on-line experimental design

Ilia Nouretdinov I.R.NOURETDINOV@RHUL.AC.UK
Royal Holloway, University of London, London, UK

Abstract

Conformal prediction is a recently developed framework of confident machine learning with guaranteed validity properties for prediction sets. In this work we study its usage in reversed version of the traditional machine learning problem: prediction of objects which can have a given label, instead of usual prediction of labels by objects. It is meant that the label reflect some desired property of the object. For this kind of task, the conformal prediction framework can provide a prediction set that is a set of objects that are likely to have the label. Based on this, we create an on-line protocol of experimental design. It includes a choice criterion based on conformal output, and elements of transfer learning in order to keep the validity properties in on-line regime.

Keywords: Confident classification, conformal prediction, experimental design, transfer learning.

1. Introduction

This work is motivated by an experimental design problem which is likely to appear in such areas as drug design. Assume that there is a set of instances (e.g. chemical compounds) and the task is to find an item with a desired property. For any of the items, this can be done by experimental validation but this is costly. The success (reward) is the percentage of experiments which were successful in the sense that the selected object had really shown to have the desired property. On-line setting assumes that after selecting an instance, a natural experiment makes its label known for further research.

A trivial way is to select all the instances for experimental validation randomly. We consider this as a 'baseline': a non-random choice strategy is successfull if it leads to a higher percentage of success.

If no labels are open initially, there is no other choice than to make at least few first experiments by a purely random selection. So we divide the learning into two parts. On the first phase some number of instances is selected randomly. On the second phase each instance is selected intentionally, based on the results of machine learning. The strong separation between two stages because the data selected on the first phase is fairly representative that will help with keeping validity further.

The contribution of this paper is applying reverse conformal framework for the aim of this choice. Some review of conformal active learning was presented in (7) but there were no validity guarantees for on-line version when new validated examples are added to the data. We aim to keep validity properties guaranteed by conformal prediction, and at the same time to present new ways of using conformal output for efficient decisions.

The key ideas will be explained in background Section 2 where we explain the idea of reverse (object-by-label) conformal approach. Two resulting algorithms are presented in Section 3. They are different in the approach to selection of an instance during the second phase of learning. Section 4 is modelling such kind of on-line experimental design on the base of a toy problem where the natural experiments are retrospectively simulated by a formal 'opening' of the labels for the learning algorithm. The task is the search for 'edible' mushrooms, based on the Mushroom data set (6) from UCI repository. As an experimental goal, we try to increase percentage of success that is measured by the percentage of edible mushroom find in 100 first 'experiments'. The improvement is being done in the following ways. First, by looking for the right balance between random and active phases of learning. Second, by applying a specially designed criterion of choice for the second phase. We finish with conclusion Section 5 discussing directions of the future work.

2. Background

2.1. Conformal prediction

The task of machine learning is to predict a label for a new (or a testing) example x_{l+1} from a given training set of feature vectors $x_1, x_2, \dots, x_l \in X$ supplied with labels $y_1, y_2, \dots, y_l \in Y$. The conformal prediction technique introduced in (1) and had many applications and extensions later. It allows to make a valid confident prediction. In conformal prediction approach for supervised learning a (feature vector, label) pair (x_i, y_i) is understood as a whole object z_i.

The core detail of conformal predictor for a non-conformity measure (NCM) A that is a measure of information distance an object z, which is usually a labelled feature vector, and a set U of objects of the same nature. In other way it can be said that NCM estimates relative typicalness of the objects $z_1, \dots, z_{l+1}$ with respect to each other:

$$\alpha_i = A\left(z_i, \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_{l+1}\}\right).$$

In the case of supervised learning (classification or regression), a p-value is assigned to each possible hypothesis $y \in Y$ about the label of the new object $x_{l+1} \in X$:

$$p(y) = p(z_1, \dots, z_l, (x_{l+1}, y)).$$

calculated as

$$p = \frac{card\{i = 1, \dots, l+1 : \alpha_i \geq \alpha_{l+1}\}}{l+1}$$

The predictions follow two useful properties: validity and efficiency. The validity property states that if the sequence of examples $z_1, \dots, z_{l+1}$ are really generated by an i.i.d. (power) distribution then for any significance level ε, the probability that $p < \varepsilon$ is at most ε.

In supervised case, the sequence $z_{l+1} = (x_{l+1}, y_{l+1})$ is known partially, the task is to predict y_{l+1} by x_{l+1}. The prediction set $R^\varepsilon \subset Y$ is the set of y s.t.

$$p(y) = p\left(z_1, \dots, z_l, (x_{l+1}, y)\right) > \varepsilon.$$

Validity implies that this R^ε covers the true label y_{l+1} with probability at least $1-\varepsilon$. This makes the prediction set reliable: probability of error is limited by ε, if error is understood as a true label left outside the prediction set. Reliability of p-values and of the prediction sets will also allow as to apply for the decisions.

The supervised conformal predictor outputs the prediction set for a given significance level ε. The smaller it is the more efficient (informative) the prediction is. The criteria of efficiency of conformal anomaly detection is discussed in detail in (5), as well as efficiency of supervised learning in (4). Validity has the first priority, and efficiency should be increased only as far as it does not affect the validity. Further we will explain what is meant by efficiency for our task of on-line experimental design.

2.2. Reverse (object-by-label) conformal prediction

Assume now that the problem is the opposite: to guess which object x should have a desired label y. In this case the roles of x and y are swapped and the prediction set is a subset of the object space X. Further in this work we will call it x-prediction set to distinguish it from the usual (y-) prediction set. For this problem we need to select a testing set S which may the whole space X or its subset.

Generally, the x-prediction set $R_h^\varepsilon \subset S \subset X$ is the set of $x \in S$ s.t.

$$p_h(x) = p\left(z_1, \ldots, z_l, (x, h)\right) > \varepsilon.$$

The validity property for x-prediction set may be understood in the following way. Assume that we are looking for an example labelled $y = h$ within the testing set S. If $x \in S$ does really have this property (label $y = h$), then it is covered by R_h^ε with probability at least $1-\varepsilon$.

Efficiency of the prediction can be measured by small size of the prediction set for given significance level ε, or by average p-value on S. Some sort of a similar task was considered in the work (2), but it was a partially supervised problem: the training set included only the examples with the label $y = 1$.

2.3. Conformal prediction for transfer learning

Conformal prediction for transfer learning is developed in (3). It is needed to save validity properties for some deviations from i.i.d. assumption.

We need to use in the case when the training set is extended by examples chosen in a not completely random way. One can say that there are two training sets instead of one: main (target) set T and addition (source) set S.

The learning is done in the following way. Let $T = \{z_1, \ldots, z_t\}$, $S = \{z_{t+1}, \ldots, z_l\}$; non-conformity scores can be defined as before:

$$\alpha_i = A\left(z_i, \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{l+1}\}\right).$$

However calculation of p-values is changed:

$$p = \frac{card\{i = 1, 2, \ldots, t, l+1 : \alpha_i \geq \alpha_{l+1}\}}{t+1}.$$

This allows to save validity property although S has a deviation from T in the generating distribution.

3. Methodology

Let the data set be presented as a set of examples $z_i = (x_i, y_i)$ $(i = 1, \dots, N)$ as edges and connections (i, j) as nodes. In this work we assume the case of binary labels $y_i \in \{0, 1\}$. At each time moment, some of the labels are available for the learner, while the others are hidden.

Denote as K_0 the set of i such that y_i is known before the active phase of learning starts. We can assume that this data set is the result of an initial experimental phase where a given amount of the objects were selected fairly randomly. Therefore it can be used as the target in terms of transfer learning.

At each of the following time moments $(t = 1, 2, \dots, B)$ of the second phase, the learner is allowed to open one more label y_j of the example $j = j_t$, so $K_t = K_{t-1} \cup \{j_t\}$. The number j can be chosen with some restriction: $j \in C_t$ where

$$C_t = F(K_{t-1})$$

meaning that the possibility of choice of the next object for experimental testing is somehow limited by the set of already tested objects. Due to violation of i.i.d. assumption by intentional choice of the instances, the labelled examples obtained in these experiments can be used in further learning, but only as a source of transfer. Effectiveness of such learning is measured by the number of discoveries of j_t such that $y_{j_t} = 1$.

Algorithm 1 Selection by largest p-value

```
$A$ and $B$ are lengths of random and active phases
start with a randomly chosen $K_0 \subset \{1, \dots, N\}$ of size $A$
FOR $t := 1$ TO $B$
*** The suggested rule of choice for $j_t$ ***
create the candidate set: $C_t = F(K_{t-1})$
FOR $c \in C_t$
train CP on $K_0$ with transfer from $K_{t-1} \setminus K_0$
test on $x_c$
assign p-value $p_c$ to the hypothesis that $y_c = 1$
ENDFOR
*** Select the candidate with the largest p-value ***
$j_t = \arg\max_{c \in C_t} p_c$
$K_t = K_{t-1} \cup \{j_t\}$.
*** Now $y_{j_t}$ becomes open ***
ENDFOR
```

In Algorithm 1 the choice criterion is the largest p-value assigned to the positive hypothesis. This can be understood just as selection of 'the most likely' object. The disadvantage of this method is that p-value include large element of randomness. The alternative approach is inspired by (2) where the size of (x-)prediction set was used as a measure of prediction success. Let us not just compare p-values to a significance level (threshold), but also use the small size of x-prediction set as another criterion of choice. This imitation of future prediction may be done in assumption that the training set (its source part) is extended with the new example assigned the label 0. In other words, we measure the reward in terms of making the prediction set narrow, assuming that the choice of the new example was not successful in direct reward. This is reflected in Algorithm 2.

Algorithm 2 Selection by largest reward

```
start with $K_0 \subset \{1, \dots, N\}$
INPUT significance level $\varepsilon$
FOR $t := 1$ TO $B$
*** The suggested rule of choice for $j_t$ ***
create the candidate set: $C_t = F(K_{t-1})$
create a randomly selected testing set: $S \subset \{1, \dots, N\} \setminus (K_{t-1} \cup C_t)$
FOR $c \in C_t$
train CP on $K_0$ with transfer from $K_{t-1} \setminus K_0$
test on $x_c$
assign p-value $p_c$ to the hypothesis that $y_c = 1$
END FOR
create the short list: $C'_t = \{c \in C_t : p_c > \varepsilon\}\}$
FOR $c \in C'_t$
train CP on $K_0$ with transfer from $(K_{t-1} \cup \{c\}) \setminus K_0$ (with assumed $y_c = 0$)
test on $S$
measure the reward $r_c = card\{s \in S : p_s < \epsilon\}$ where $p_s$ is p-value assigned to the hypothesis $y_s = 1$
ENDFOR
***Select the candidate with the largest reward amongst ones with high p-values***
$j_t = \arg\max_{c \in C'_t} r_c$
$K_t = K_{t-1} \cup \{j_t\}$.
*** Now $y_{j_t}$ becomes open ***
ENDFOR
```

4. Application

4.1. Data processing and modelling

As a toy example, we use the Mushroom data set from UCI repository (6). The data set contains 8,124 instances with 22 discrete attributes and 2 classes (whether a mushroom is edible or poisonous). 4,208 instances blog to the positive (edible) class.

In almost all cases the mushrooms different in 1-2 attributes are either both edible or both poisonous. Therefore 1-Nearest-Neighbours algorithm is enough for our modelling aims. The non-conformity score of an object is defined as Hamming distance to the nearest same class object divided by Hamming distance to the nearest other class object.

We assume that the number of possible experiments is 100, which is the sum of random and active phases of the learning:

$$A + B = 100.$$

We also assume that on any step of the second phase we can not select any arbitrary instance from the data set, but the choice is limited with some constraints. This models

ϕ	A	B	Reward	ϕ	A	B	Reward	ϕ	A	B	Reward
1	2	98	62.72	0.2	2	98	29.18	0.1	2	98	14.88
1	6	94	85.80	0.2	6	94	41.77	0.1	6	94	22.94
1	10	90	88.58	0.2	10	90	45.29	0.1	10	90	25.16
1	12	88	88.80	0.2	12	88	45.89(best)	0.1	12	88	25.46
1	14	86	88.85	0.2	14	86	45.01	0.1	14	86	26.38
1	16	84	88.97(best)	0.2	16	84	43.71	0.1	16	84	26.66(best)
1	18	82	88.75	0.2	18	82	44.39	0.1	18	82	26.50
1	20	80	88.52	0.2	20	80	43.06	0.1	20	80	25.36
1	30	70	85.39	0.2	30	70	38.08	0.1	30	70	23.09
1	40	60	81.28	0.2	40	60	34.52	0.1	40	60	20.42
1	50	50	76.34	0.2	50	50	30.74	0.1	50	50	18.09
1	60	40	71.44	0.2	60	40	28.19	0.1	60	40	15.89
1	70	30	66.70	0.2	70	30	26.31	0.1	70	30	14.52
1	80	20	61.75	0.2	80	20	23.06	0.1	80	20	12.87
1	90	10	57.20	0.2	90	10	20.47	0.1	90	10	11.03
1	100	0	52.20	0.2	100	0	17.80	0.1	100	0	9.62

Table 1: Results for Algorithm 1

a possible situation when generating new compounds is made step-by-step, by chemical reactions applied to already existing compounds. According to that, we assume as well that the space of choice may change from step to step depending on what labels are already opened. In our setting we assume that the object's label can be opened only if the position of the object in the original data base follows a position of already opened label. Here 1 is assumed to follow the last position (that is 8124 if the full dataset is used). For example: if the labels are already opened for the instances 20, 33, 34, 8124 then then the next label for opening can be chosen only from 1, 21, 35 only. The motivation of this statement is to model a situation when the influence of a mutation/reaction (from 20 to 21) on the features is very indirect.

4.2. Results

The results for the Algorithm 1 are presented in Table 1. In the table $\phi < 1$ means that the positive (edible) class is reduced to this percentage of its original size, while the poisonous class remained the same. This is done to make the searching problem harder. The experimental results are averaged over 100 random seeds.

The bottom line of Table 1 corresponds to the 'baseline': all the examples are chosen randomly, so the reward corresponds to the percentage of positives in the data selection that is full data set for $\phi = 1$ and its imbalanced part for $\phi < 1$. In the other lines A examples are selected randomly B steps are done in random way. In all of the experiments the reward is better than the completely random choice. The best length of the first phase is 16 (for $\phi = 1$ and $\phi = 0.1$) and 12 (for $\phi = 0.2$). Surprisingly, this optimum does not essentially depend on the level of imbalance in the data.

ϕ	A	B	Algorithm 1	Algorithm 2 (Sec. 2)		
			(Sec. 1)	$\epsilon = 0.05$	$\epsilon = 0.01$	$\epsilon = 0.001$
0.1	2	98	14.88	15.39	15.39	15.39
0.1	6	94	22.94	26.50	26.50	26.50
0.1	10	90	25.16	31.38	31.38	31.38
0.1	12	88	25.46	33.15	33.15	33.15
0.1	14	86	26.38	35.47	35.47	35.47
0.1	16	84	26.66(best)	36.18 (best)	36.18	36.18
0.1	18	82	26.50	15.37	36.40 (best)	36.43 (best)
0.1	20	80	25.36	22.62	36.07	36.07
0.1	30	70	23.09	20.64	34.93	34.93
0.1	40	60	20.42	18.92	32.45	32.45
0.1	50	50	18.09	16.67	28.72	28.72
0.1	60	40	15.89	15.20	25.10	25.10
0.1	70	30	14.52	13.49	21.21	21.21
0.1	80	20	12.87	12.56	17.21	17.21
0.1	90	10	11.03	10.85	13.26	13.26
0.1	100	0	9.62			

Table 2: Results for Algorithms 1 and 2

The comparison of Algorithms 1 and 2 for the most interesting case is presented in Table 2. The size of randomly selected testing set is 1% of the data, ϵ is set to 5%. In most of the cases the second algorithms gives essential improvement. The exceptions (with $A = 18$, $A = 20$) may be caused by closeness of $1/A$ to ϵ, so using $\epsilon = 0.01$ is preferable. Further decreasing of ϵ lead to slight improvement at the cost of high computational load.

5. Conclusion

This paper have shown how the experimental design can be done on the based on the conformal prediction, and how the validity of conformal prediction can be saved by means of transfer learning. Although conformal prediction is based on i.i.d. assumption and therefore requires some part of the data for the experiments to be selected fairly randomly, this part is usually not a large one ($12 - 16\%$) as shown in the experiments.

However, this is a toy example, and there are question for future research. The main of them is how to increase the achieved quality further. Application of Algorithm 2 shows a possible direction of improvement, requiring more studies. The results are promising although the method is more time-consuming.

6. Acknowledgments

This work was supported 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; by European Union grant

671555 (”ExCAPE”); EPSRC grant EP/K033344/1 (”Mining the Network Behaviour of Bots”); and by the National Natural Science Foundation of China (No.61128003) grant.

We are grateful to Lars Carlsson and Claus Bentzen for the initial motivation and discussion.

References

[1] Vovk, V., Gammerman, A., Shafer, G. Algorithmic Learning in a Random World. Springer, 2005

[2] Ilia Nouretdinov, Alex Gammerman, Yanjun Qi, Judith Klein-Seetharaman Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, p.311, 2012.

[3] Shuang Zhou, Evgueni N. Smirnov, Haitham Bou Ammar, Ralf Peeters. Conformity-Based Transfer AdaBoost Algorithm Artificial Intelligence Applications and Innovations. Volume 412 of the series IFIP Advances in Information and Communication Technology pp 401-410, 2013.

[4] Vovk, V., Fedorova, V., Nouretdinov, I., Gammerman, A. Criteria of efficiency for conformal prediction. 2014 On-line COMPRESSION Modelling Project (New Series), 19 p.

[5] Smith, J., Nouretdinov, I., Craddock, R., Offer, C., Gammerman, A. Anomaly Detection of Trajectories with Kernel Density Estimation by Conformal Prediction. Artificial Intelligence Applications and Innovations: AIAI2014 Workshops. Rhodes, Greece: Springer, pp. 271–280.

[6] Mushroom Data Set. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Mushroom

[7] Balasubramanian, V.N., Chakraborty, S., Ho, S.-S.,Wechsler, H., Panchanathan, S. Active learning. In: Conformal Prediction for Reliable Machine Learning Theory, Adaptations and Applications Edited by: Vineeth Balasubramanian, Shen-Shyang Ho and Vladimir Vovk. Elsevier, 2014, Pages 49-70.