Conformal Prediction under Probabilistic Input

Ilia Nouretdinov

Computer Learning Research Centre,
Royal Holloway University of London

Abstract. In this paper we discuss a possible approach to confident prediction from data containing missing values presented in a probabilistic form. To achieve this we revise and generalize the notion of credibility known in the theory of conformal prediction.

1 Introduction

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 $x_1, x_2, \ldots, x_l$ supplied with labels $y_1, y_2, \ldots, y_l$. The conformal prediction technique introduced in [1] and had many applications and extensions later. It allows to make a valid confident prediction.

Originally it was introduced for supervised machine learning problem with clear data structure. But in many practical problems data representation may be complex and combine multiple sorts of information. The conformal prediction was extended in previous works. In [6], it was *semi-supervised learning* when only some examples are presented with labels. In [7] training labels were available only for one of two classes. In [8] an unsupervised learning problem of *anomaly detection* was considered. Another kind of the task is Vapnik's Learning under privileged information [4] that can be interpreted as having missing values in testing examples. A conformal approach to it was made in the work [5].

The direction presented here is probabilistic representation of feature vectors or labels. Assume that there is kind of a priori distribution on features and/or labels. For example it is concentrated at one value when a feature is presented, it is uniform when it is completely missing, and other distributions are applicable when it is known partially or hypothetically. This means neither to try to exclude examples with missing values nor to fill them in a unique way.

An approach to this task is based on the notion of credibility that appears in the standard (supervised) conformal prediction. Unlike the confidence assigned to a likely hypothesis about the new example's label, the credibility answers the question whether any of these hypotheses is true at all. So the credibility is a characteristic of an unfinished data sequence, that includes a new example without its label. This can be naturally extended to the task when some part of training information is missing.

As an area of application needed for an illustration of the proposed method, we take LED data set from UCI repository [2], because a priori distribution on the values has a clear sense for these data.

L. Iliadis et al. (Eds.): AIAI 2014 Workshops, IFIP AICT 437, pp. 291–300, 2014.

2 Machine Learning Background

2.1 Conformal Prediction

Let us remind the properties of conformal prediction (in the case of classification) according to [1].

Assume that each data example z_i consists of x_i that is an m-dimensional vector $x_i = (x_{i1}, \ldots, x_{im})$ and a label y_i that is an element of a finite set Y.

Conformal predictor in supervised case assigns p-value (the value of a test for randomness) to a data sequence

$$p(y) = p((x_1, y_1), \ldots, (x_l, y_l); (x_{l+1}, y))$$

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

where $x_1, \ldots, x_l \in X$ are feature vectors of training examples with known classifications $y_1, \ldots, y_l \in Y$, x_{l+1} is a new or testing example with a *hypothetical* label y, and

$$\alpha_i = A\left(\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y)\}, (x_i, y_i)\right)$$

for a *nonconformity measure* A that is a strangeness function of a set of labeled feature vectors and one of its elements.

The plan is to check each possible hypothesis about the label of a new example, and to the label of new example would conform the assumption of exchangeability, or with which label the example 'fits well' into the training set? The prediction set consists of satisfactory hypotheses y such that $p(y)$ exceeds a *significance level* γ. The calculations of prediction regions are based on a special function called *nonconformity measure (NCM)* that reflects how strange an example is with respect to others. Then p-value is assigned to each y.

There are two ways to present the results. One of them is the *prediction set*: a list of y which meet this confidence requirement $p(y) \geq \gamma$. The *validity* property implies that the probability of error is at most $1 - \gamma$ whenever the i.i.d. assumption is true. Here an error means true value of y_n being outside the prediction set.

Alternatively we can provide the prediction of a new label together with measures of its individual *confidence*. The correspondence between two types of output is that the confidence is the highest confidence level at which the prediction region consists of (at most) one value. In terms of p-values assigned to different labels, the confidence is a complement to 1 of the second highest p-value.

An individual prediction is also naturally completed with *credibility* that is the first highest p-value. If the credibility is low this means that any existing hypothesis about the label of the new object is unlikely. In other words, the new object itself is not credible enough as a continuation of the data sequence, and this could be said before its label is known. So it can be understood as dealing with an unknown testing label.

Our aim is to extend this idea, dealing other sort of incomplete information in analogous way.

2.2 Standard Credibility

In this work we call credibility a measure of conformity of an incomplete data sequence.

Originally it was applied to the data sequences of the following type:

$$(x_1,y_1),\ldots,(x_l,y_l),x_{l+1}.$$

with y_{l+1} missing.

The credibility is obtained by maximization conformal p-values over all its possible completions:

$$p_{cred}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1})$$

$$= \max_{y_{l+1}\in Y} p((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y_{l+1})).$$

The validity property of conformal prediction regions can be easily extended to the credibility. If a data sequence $(x_1,y_1),\ldots,(x_{l+1},y_{l+1})$ is generated by $P = P_1^{l+1}$ where P_1 is a distribution on $X \times \{0,1\}$, then

$$P\{p_{cred}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1}) \le \gamma\} \le \gamma$$

for any $\gamma \in (0,1)$.

In this form it was assumed that the incomplete sequence is obtained from the complete one by forgetting y_{l+1}. In other words it could be said that the incomplete sequence $(x_1,y_1),\ldots,(x_l,y_l),x_{l+1}$ is generated by $P^l \times P_X$ where P_X is the marginal distribution of the feature vector averaged over Y.

2.3 Extensions of Credibility

As we have seen, the standard credibility is the p-value (test for randomness) assigned to an incomplete sequence of examples. Incompleteness means there that the label of the last example is totally missing.

Sometimes a similar aproach can be applied to other kinds of missing values. A close problem is having an unknown feature (not a label) of a new (testing) example. This task is equivalent to learning under privileged (additional) information framework formulated in [4]. The conformal approach of this task was developed in the work [5]. An analogue of credibility was assigned to the sequence

$$(x_1,x_1^*,y_1),\ldots,(x_l,x_l^*,y_l),(x_{l+1},y)$$

with x_{l+1}^* unknown. The feature x^* was called privileged bacause it is available for the training examples.

Next step might be related to missing values in training examples. But a straightforward approach to this task (maximizing p-value over possible fillings of the gap) is not effective because the conformal predictor concentrates on the conformity of the testing example without checking training examples for strangeness. Therefore we would like to consider missing values as distributions.

3 Conformal Approach for Probabilistic Input

For convenience of presentation, in this section we will start form the case of unclear information about binary labels y_i presented in a probabilistic form of a priori distribution. Then we will show how to apply it in a more general case.

3.1 Task and Assumptions

Suppose that $Y = \{0,1\}$, but some information about $y_1,\dots,y_l$ is missing. However, for each $i = 1,\dots,l$ we know that p_i has a meaning of probability that $y_i = 1$. As for y_{l+1}, we assume that it is known as a hypothesis according to the conformal prediction procedure.

How to state this task in a well-defined way and what would be a proper analogue of the i.i.d. assumption in this case?

A mechanism should generate both the 'true' data sequence (including hidden values of y_i) and the 'visible' one (with probabilistic values p_i). This means that the triple (x_i, y_i, p_i) is generated simultaneously. But some agreement between p_i and y_i is also needed so that probabilistic values p_i make sense as probabilities.

To define this formally, assume that P_1 is a distribution on $X \times (0,1)$ and Θ is the uniform distribution on (0,1). First, $P = (P_1 \times \Theta)^{l+1}$ generates

$$(x_1, p_1, \theta_1), \dots, (x_{l+1}, p_{l+1}, \theta_{l+1}).$$

Setting $y_i = 1$ if $p_i < \theta_i$ and $y_i = 0$ otherwise, we can also say that P^* generates a sequence of triples

$$(x_1, p_1, y_1), (x_2, p_2, y_2), \dots, (x_{l+1}, p_{l+1}, y_{l+1})$$

where p_i is 'visible' label and y_i is the 'hidden' one, y_i is stochastically obtained from p_i.

3.2 Special Credibility

In order to make a conformal prediction of y_{l+1} for x_{l+1} we need to consider different hypotheses about it. When a hypothesis is chosen, we work with 'visible' labels $p_1,\dots,p_l$ for training examples and for a 'hidden' value y_{l+1} for the new one. Thus the task is to assign a valid credibility value for a sequence $(x_1, p_1), \dots, (x_l, p_l); (x_{l+1}, y_{l+1})$.

Fix a parameter $s > 0$ called *allowance* which is a trade-off between testing the hypothetical new label with respect to a version of the training data set, and testing the training data set with respect to a priori distribution on missing values.

Suppose that Q is the conditional distribution of $p((x_1,y_1),\dots,(x_l,y_l),(x_{l+1},y_{l+1}))$ given $(p_1,\dots,p_l)$ and y_{l+1}, $q_0 = q_0(p_1,\dots,p_l)$ is the smallest q such that

$$Q\{p > q | p_1,\dots,p_l; y_{l+1}\} \le s$$

and

$$p_{cred} = p_{cred}\left((x_1,p_1),\dots,(x_l,p_l);(x_{l+1},y_{l+1})\right) = q_0(p_1,\dots,p_l;y_{l+1}) + s.$$

Proposition 1. *Assume tgat* $(x_1,p_1),\ldots,(x_l,p_l)$ *and* $(x_{l+1},y_{l+1}$ *are generated by the mechanism described in Section 3.1 and* p_{cred} *us calculated as in Section 3.2, then*

$$P\{p_{cred} \leq \gamma\} \leq \gamma$$

for any $\gamma \in (0,1)$.

Proof: Recall that $p = p(y_1,\ldots,y_l) > q = q(p_1,\ldots,p_l)$ with probability at most s for any given $p_1,\ldots,p_l$. On the other hand, p is valid as a standard conformal predictor's output thus $p \leq \gamma - s$ with overall probability at most $\gamma - s$. Therefore $\gamma - s < p < q$ with probability at least $1 - s - (\gamma - s) = 1 - \gamma$ and probability that $q + s < \gamma$ is bounded by γ. □

3.3 Missing Values in Features

For convenience of presentation we earlier assumed that the labels $y_1,\ldots,y_l$ are given in probabilistic form, although this can be extended to the objects $x_1,\ldots,x_l$ as well.

So let us now assume that P^* generates (H_i,x_i,y_i) where 'visible' H_i is a distribution on X, while 'hidden' $x_i \in X$ is randomly generated by H_i.

If x_i is known clearly, this means that H_i is a distribution concentrated at one point. Otherwise H_i can be understood as an a priori distribution on its missing values. If X is discrete, then H_i can be presented in a vector form.

The extended credibility is defined by analogy. Suppose that Q is the conditional distribution of $p((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y_{l+1}))$ given $(H_1,\ldots,H_l,H_{l+1})$ and a fixed y_{l+1}, q_0 is the smallest q such that

$$Q\{p > q | H_1,\ldots,H_{l+1};y_{l+1}\} \leq s$$

and

$$p_{cred}\,((H_1,y_1),\ldots,(H_l,y_l),(H_{l+1},y_{l+1})) = q_0 + s.$$

Obviously an analogue of Proposition 1 is also true in this case.

3.4 Efficient Approximation

To find q_0 exactly one has to know the condition distribution of p given 'visible' data. For the aims of computational efficiency this distribution can be replaced with an empirical one, using Monte-Carlo approximation. Let $H_1 \times H_2 \times \ldots \times H_{l+1}$ generate a large amount of vectors $(x_1,\ldots,x_{l+1})$ and calculate conformal p-value for each of them. Then we will get an empirical distribution of p that allows to estimate q_0 by sorting these p-values and taking one with corresponding rank. An example will be given in Section 4.3.

4 Experiments

For the experiments we use benchmark LED data sets generated by a program from the UCI repository[2].

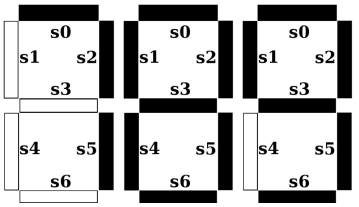

Fig. 1. Canonical images of 7,8,9 in LED data

LED means light emitting diode electronic display. A LED image of a digits has of 7 binary features (pixels). The task is to predict a digit from an image in the seven-segment display. Each of digits $0,1,\ldots,9$ has a canonical image that should normally represent it. Few examples are presented on Fig.1.

Assume now that visible displays can contain mistakes. Each pixel can occasionally show 'on' instead of 'off' or vice versa with probability p_0. For our example we assume that $p_0 = 0.1$ although normally it is much less. The data generating program first randomly selects a canonically represented digit then each of the attributes is inverted with a probability of noise p_0 and the noisy example is added to the data set.

In the work [3] the conformal approach was applied to LED data in its standard supervised form. Now we make some changes in the data statement. First, the probability p_0 itself is known for us. T his means that all values in the training set are probabilistic ones. When we see that a pixel is 'on' this in fact means that it is on with probability $1-p_0$ and 'off' with p_0, and vice versa. Second, in the testing examples there are no mistakes (as if $p_0 = 0$). The task is to classify a testing example with full information after training on the examples with probabilistic information. It is assumed that the canonical representations are not available for the learner, who has to make predictions based only on the examples with possible mistakes as they are presented in the data.

For experiments we generate some amount of LED digits. The number and distribution (frequency) of labels (0,1,2,...,9) is not restricted, we borrow it from well-known USPS (US Postal Service) benchmark data set in order to have imbalanced classes. Size of the classes is shown in Table 1.

Table 1. Size of different training classes

Class label (digit)	0	1	2	3	4	5	6	7	8	9	Total
Number of examples	359	264	198	166	200	160	170	147	166	177	2007

For a training example, given a label, we take its canonical LED image and make an error in each of the feature with probability γ. In the most of experiments $\gamma = 0.1$ unless stated another.

Testing examples are not probabilistic by the task, so in principle we can make predictions on $2^7 = 128$ possible images. This number includes 10 canonical images of digits.

Later we will consider two types of testing set. A proper one is generated with the same distribution on classes as the training set and therefore contains only canonical images. An auxiliary testing set contains all the possible images.

4.1 Nonconformity Measure

For convenience we use one of the simplest NCM that can be applied. As the space is discrete, NCM of an example with respect to the set of another ones is defined as the number of 'zero distance other class neighbors', i.e. number of examples in the set that have the same features but in fact belong to another class:

$$\alpha_i = A\left(\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\},(x_i,y_i)\right) = card\{j : x_i = x_j, y_i \neq y_j\}.$$

4.2 Probabilistic Values of the Features

We apply our approach in its form mentioned in Section 3.3. All the features of X are binary and we assumed that the mistakes in features are done independently of each other. The connection between a 'hidden' vector

$$x = (x(1),\ldots,x(7)) \in X = \{0,1\}^7$$

and a corresponding 'visible' distribution H is the following. H is a distribution on X such that:

- $x(1),\ldots,x(7)$ are H-independent on each other;
- for each $j = 1,\ldots,7$, $H\{x(j) = 1\}$ is either $1-\gamma$ or γ;
- a mistake $x(j)$ in a feature $x(j)$ is done with probability γ;
- if there is no mistake in $x(j)$ then $H\{x(j) = 1\} = 1-\gamma$ if $x(j) = 1$, γ if $x(j) = 0$;
- if there is a mistake in $x(j)$ then $H\{x(j) = 1\} = \gamma$ if $x(j) = 1$, $1-\gamma$ if $x(j) = 0$.

This means that in the 'visible' features vectors all the features are probabilistic. Each of the features is either *1 with probability* $1-\gamma$*, 0 with probability* γ or *1 with probability* γ*, 0 with probability* $1-\gamma$.

4.3 Other Details

Following 3.2 we set the 'allowance' coefficient to $s = 0.01$. Following the note 3.4 we avoid scanning all possible combinations by calculating p-values as Monte-Carlo approximations. the number of trials is 1000. Further we will see that this approximation does not affect validity properties.

Summarizing, there were 1000 trials (i.e. random filling of the missing values), and consider as the approximate credibility p_{cred} the 10-th largest of these p-values plus the allowance $s = 0.01$.

5 Results

Remind that p_{cred} finally is the p-value assigned to a new example (x_{l+1}, y_{l+1}).

To check the validity we wish to check what p-value is assigned to the true hypothesis about y_{l+1}. The correspoding p_{cred} is called p_{true}.

If y_{l+1} is unknown then each possible hypothesis about its value should be checked and assigned a p-value. As well as in the standard conformal predictor, the *prediction* is the hypothesis with the largest p-value and *confidence* in it is 1 minus the second largest p-value.

5.1 Validity

According to our problem statement, the validity is checked on testing examples that do not contain uncertainty and have the same distribution as the training examples *before* introducing mistakes. Therefore, each of the testing examples is one of ten digits ($y \in \{0,1,\ldots,9\}$) presented with its canonical image x. In order to satisfy i.i.d. assumption with training set, the distribution of ten types also corresponds to one from USPS data.

The corresponding validity plot is presented on Fig.2. It show that the probability of error (true value being outside the prediction set) does not exceed the selected significance level, for example:

$$P\{p_{true} \leq 0.16\} = 0.08;$$

$$P\{p_{true} \leq 0.27\} = 0.17.$$

The validity is satisfied with some excess. The same effect is known for the standard credibility and for LUPI due to involving incomplete information into the data.

5.2 Confidence

Recall that the testing set consists only of canonical images, so there are only 10 possible different configurations.

Individual confidences for them can be seen on Fig. 5.2 (boxed items), average value is 0.87. The smallest of these confidences is 0.79 assigned to the digit 7, because this digit is mixable with 1 (Hamming distance between them is the smallest) and relatively rare in the training set.

For comparison we also included confidence values that would be assigned to all 128 possible pixel combinations (auxiliary testing set) and they are much lower (0.18 in average).

The more indefinite the data are the smaller is the achieved level of confidence. For example, if we increase the probability of mistake from $p_0 = 0.1$ to $p_0 = 0.2$ then the figures of average confidence falls down to 0.55 and 0.11 respectively.

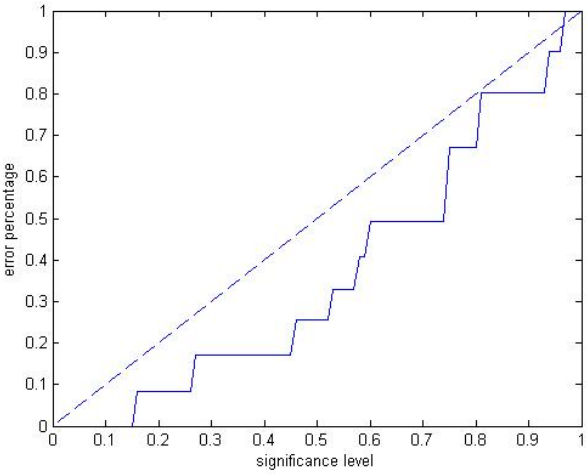

Fig. 2. Validity plot

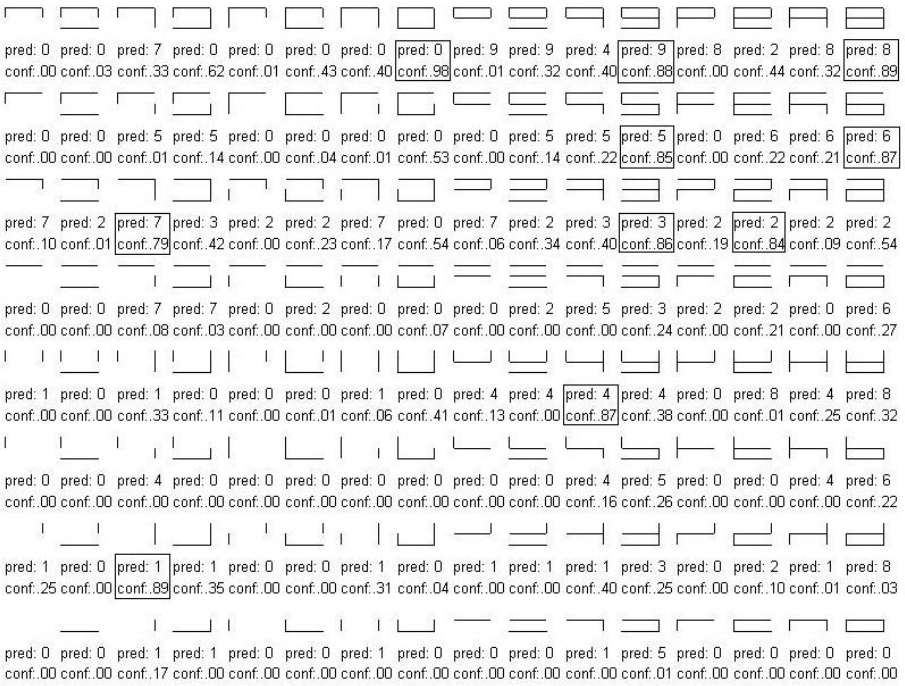

Fig. 3. Predictions for all possible pixel combinations. Predictions for the canonical images are put in boxes.

6 Conclusion

In this work we formulated an approach to get confident prediction from the data with missing values (or labels) presented in a probabilistic form. Probabilistic input means that there is an a priori distribution on possible filling of these missing values.

The advantages of conformal approach for this task are not ignoring examples with incomplete information, and on the other hand not wasting time attempting to restore the missing values.

The missing features are taken as a priori distributions on their possible values. This is an analogue of Bayesian distribution on a parameter of a statistical model. So we can expect as well that it might be assumed in other practical problems with incomplete information.

Acknowledgments. This work was supported by EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots"); by Thales grant ("Development of automated methods for detection of anomalous behaviour"); by the National Natural Science Foundation of China (No.61128003) grant; and by grant 'Development of New Venn Prediction Methods for Osteoporosis Risk Assessment' from the Cyprus Research Promotion Foundation.

We are grateful to Judith Klein-Seetharaman and Alex Gammerman for motivating discussions.

References

1. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005)
2. LED Display Domain Data Set, `http://archive.ics.uci.edu/ml/datasets/LED+Display+Domain`
3. Fedorova, V., Gammerman, A., Nouretdinov, I., Vovk, V.: Conformal prediction under hypergraphical models. In: Papadopoulos, H., Andreou, A.S., Iliadis, L., Maglogiannis, I. (eds.) AIAI 2013. IFIP AICT, vol. 412, pp. 371–383. Springer, Heidelberg (2013)
4. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural Netwotks 22, 544–557 (2009)
5. Yang, M., Nouretdinov, I., Luo, Z.: Learning by Conformal Predictors with Additional Information. In: Papadopoulos, H., Andreou, A.S., Iliadis, L., Maglogiannis, I. (eds.) AIAI 2013. IFIP AICT, vol. 412, pp. 394–400. Springer, Heidelberg (2013)
6. Adamskiy, D., Nouretdinov, I., Gammerman, A.: Conformal prediction in semi-supervised case. In: Post-Symposium Book 'Statistical learning and Data Science'. Chapman and Hall, Paris (2011)
7. Nouretdinov, I., Gammerman, A., Qi, Y., Klein-Seetharaman, J.: Determining Confidence of Predicted Interactions Between HIV-1 and Human Proteins Using Conformal Method. In: Pacific Symposium on Biocomputing, vol. 17, pp. 311–322 (2012)
8. Lei, J., Robins, J., Wasserman, L.: Efficient Nonparametric Conformal Pre- diction Regions. arXiv:1111.1418v1