# Learning by Conformal Predictors with Additional Information

Meng Yang, Ilia Nouretdinov, Zhiyuan Luo

# Learning by Conformal Predictors with Additional Information

Meng Yang, Ilia Nouretdinov, Zhiyuan Luo

Computer Learning Research Centre, Royal Holloway, University of London
Egham Hill, Egham, Surrey TW20 0EX, UK
*m.yang@cs.rhul.ac.uk; ilia@cs.rhul.ac.uk; zhiyuan@cs.rhul.ac.uk*

**Abstract.** In many supervised learning applications, the existence of additional information in training data is very common. Recently, Vapnik introduced a new method called LUPI which provides a learning paradigm under privileged (or additional) information. It describes the SVM+ technique to process this information in batch mode. Following this method, we apply the approach to deal with additional information by conformal predictors. An application to a medical diagnostic problem is considered and the results are reported.

*Keywords*: LUPI, additional information, conformal predictor

## 1 Introduction

In machine learning classification problems, in batch setting, we usually work with a set of training and testing examples. In a data-rich world, there often exist some "pieces" of information about the data that we can add and use it. But, this information may be available at a training stage and not for the new examples at the testing stage. For example, usually doctors try to make diagnosis using all available information, but if at the end of an investigation the diagnosis is still unclear, they may send the patient for some additional tests such as pathological reports, blood test, MRI scans, etc. This is additional or privileged information and can be used to improve the quality of training set and hence, the decision rules. However, the same additional information may not be available for new patients. The question is: can this additional information at the training stage improve the accuracy of diagnosis for the new patients? Traditional learning methods cannot use the additional information directly when it is not available in test set – it is summarised Table 1. Recently, Vapnik proposed a

**Table 1.** Data set with additional information

| Data Set | Content | | |
|---|---|---|---|
| | 'Usual' information | Additional information | Label |
| Training examples | Known | Known | Known |
| Test examples | Known | Unknown | To be predicted |

general approach to deal with this problem, known as Learning Using Privileged Information (LUPI) [10]. However, LUPI approach does not allow us to estimate confidence in the prediction. This paper extends the Conformal Predictors method [2] to include some additional information available in the training set in order to make prediction, estimate confidence of the prediction and apply it in batch and on-line mode.

## 2 Learning using privileged information

Learning using privileged information (LUPI) is a recently proposed learning paradigm and the aim is to incorporate that type of information into learning [10]. An example of privileged information, according to Vapnik, is when teachers provide students with extra knowledge which exists in explanation, comments, comparisons and so on. There is no formal definition of "privileged" information, but we shall interpret it as information that exists only in the training set.

Let's consider a sequence of examples $x$ with their labels $y$:

$$(x_1, x_1^*, y_1), (x_2, x_2^*, y_2), ..., (x_{n-1}, x_{n-1}^*, y_{n-1}), x_i \in X, x_i^* \in X^*, y_i \in Y.$$

Here $x_i \in X$ is an example $i$ that is a vector of attributes of "usual" or "available" information and $x_i^* \in X^*$ is a vector of additional (or "privileged") attributes; $y_i$ is a corresponding label.

In the classical SVM a prediction for the new example $x_n$ can be calculated by the following equation:

$$\hat{y}_n = \sum_{i=1}^{n-1} \alpha_i y_i (x_i \cdot x_n)$$

where weighting coefficients $\alpha_i$ are calculated on the basis of the examples $x_1, \ldots, x_{n-1}$ and $x_n$ is a new example from the test set. A new method, SVM+ is an extension of SVM and Lagrange multipliers $\alpha_i$ are replaced with $\alpha_i^*$ calculated from $x_1^*, \ldots, x_{n-1}^*$, while the dot product $(x_i \cdot x_n)$ is not changed to $(x_i^* \cdot x_n^*)$ because $x_n^*$ is unavailable.

## 3 Conformal approach

### 3.1 Conformal Predictors

Conformal predictor is a general learning framework to make well-calibrated predictions, and provides predictions with reliable measures of confidence. The prediction is based on the statistical $p$-value, which is derived from the strangeness (or non-conformity) measure $\alpha_i$, that indicates how "strange" a particular example is. Any strangeness measure can be used, as long as it holds the exchangeability property. Strangeness measures may be constructed from almost any

existing learning algorithms, such as Neural Networks [3], Random Forests [6] and SVMs [11]. In this paper, we consider the Nearest Centroid method [8] to derive the strangeness measure. In general, given a strangeness measure $A$, the corresponding values are computed for each hypothetic label $y \in |Y|$ as

$$\alpha_i = A(\wr(x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ...(x_n, y)\wr, z_i), i = 1, ..., n-1$$

Given a strangeness measure we can compute p-values:

$$p(y) = \frac{\#\{i = 1, ..., n : \alpha_i \geq \alpha_n\}}{n}$$

Obviously, $0 < p_y \leq 1$. The lower p-value is, the more "strange" the example is in relation to the entire training set.

### 3.2 Learning with Additional Information

Let's consider a sequence of independent and identical examples $x$ with additional information $x^*$ and their labels $y$. For the prediction of new object $x_n$, we firstly assign it an hypothetic label ($y$) and hypothetic values ($x^*$) of additional attributes and then measure how "strange" the new example is by calculating $p(y, x^*)$. The more likely the hypothetic label is, the higher extended $p$-value $p(y, x^*)$ is. However, the number of possible combinations will affect the speed of the processing.

The advantage of Conformal Predictors is its validity, which means:

$$Prob\{p(y) \leq \varepsilon\} \leq \varepsilon$$

for any $0 < \varepsilon < 1$. Therefore, our next task is how to combine a number of extended $p$-value $p(y, x^*)$ into $p(y)$ and to maintain the validity property. Since only one of the hypotheses is true, selecting the maximum extended $p$-values is the only way to hold the validity:

$$\max_{x^*} p(y, x^*) \geq p(y, x_{true}), y \in Y, x^* \in X^*$$

Thus:

$$Prob\{\max_{x^*} p(y, x^*) \leq \varepsilon\} \leq Prob\{p(y, x^*_{true}) \leq \varepsilon\}$$

So:

$$Prob\{\max_{x^*} p(y, x^*) \leq \varepsilon\} \leq \varepsilon$$

Excluding $x^*$ from it we would get a standard conformal predictor that ignores additional information. Algorithm 1 summarises the procedure:

This method could be applied both in the on-line mode and the off-line mode. In the on-line mode, the examples are presented one by one. Each time, we observe the object and predict its label. We could assume that after the prediction is done, both the label $y_i$ and the attribute value $x^*_i$ will be revealed,

**Algorithm 1** Learning With Additional Information

**Require:** training example sequence $z_1 = (x_1, x_1^*, y_1), z_2 = (x_2, x_2^*, y_2), ..., z_{n-1} = (x_{n-1}, x_{n-1}^*, y_{n-1})$
**Require:** new example $x_n$
**Require:** strangeness measure A
  **for** $y \in Y$ **do**
    **for** $x^* \in X^*$ **do**
      $z_n = (x_n, x^*, y)$
      **for** $i$ in $1, 2, ..., n$ **do**
        $\alpha_i = A(\wr z_1, z_2...., z_{i-1}, z_{i+1}, ...z_n\wr, z_i)$
      **end for**
      $p(y, x^*) = \frac{\#\{i=1,...,n:\alpha_i \geq \alpha_n\}}{n}$
    **end for**
    $p(y) = \max_{x^*} p(y, x^*)$
  **end for**

see in the following description of the on-line prediction with additional information protocol. At the $n$-th step, we have observed the previous examples $(x_1, x_1^*, y_1), ..., (x_{n-1}, x_{n-1}^*, y_{n-1})$ and new object $x_n$ and our task is to predict $y_n$ without $x_n^*$. The new example will be added to the training examples and used to generate a new rule for next prediction. On-line mode is a simple form of the slow learning from [11] where the feedback is given with a delay. In this protocol we assume that some symptoms may also come with a delay. For example, if a prediction algorithm is designed to classify whether a patient has a disease or not by some symptoms and blood test in on-line mode, but the blood test result is not available (will be given, maybe, one day later).

On-line prediction with additional information Protocol:
FOR $n = 1, 2....$;
Reality outputs $x_n \in X$;
Predictor outputs $\Gamma_n^\varepsilon \subseteq Y$ for all $\varepsilon \in (0, 1)$;
Reality outputs $x_n^* \in X^*$, $y_n \in Y$;
END FOR

## 4 Applications and Experiments

The conformal prediction method with additional information has been applied to Abdominal Pain dataset [1]. The data set consists of 6387 patient records with 9 categories of diseases and 135 symptoms [1, 4, 5]. The 9 diseases for diagnosis are: Appendicitis (APP, 844 examples), Diverticulitis (DIV, 143 examples), Perforated Peptic Ulcer (PPU, 130 examples), Non-Specific Abdominal Pain (NAP, 2835 examples), Cholecystitis (CHO, 572 examples), Intestinal Obstruction(INO, 417 examples), Pancreatitis(PAN, 96 examples), Renal Colic(RCO, 473 examples) and Dyspepsia(DYS, 877 examples).

Each symptom has two values, 1 and 0: either the patient has the symptom or not. For each disease group, experts suggest a sequence of symptoms which

are more relevant for its diagnosis. Suppose that some of these symptoms are known for the collected training data but are unknown for a testing example, then they play the role of privileged information in this paper.

If we now choose, for example, the Nearest Centroid algorithm as an underlying algorithm to derive the corresponding strangeness measure, by using the ratio of distances as a strangeness measure:

$$\alpha_i = \frac{\{D(x_i, \mu_y)|y_i = y\}}{\min\{D(x_i, \mu_i)|y_i \neq y\}}$$

where $D$ is the Euclidean distance measure and $\mu_i$ is the centroid (the averaged example) of the class $i$. Then, we can label a new example the same way as the examples of the nearest class.

**Table 2.** Single prediction by Conformal Predictor on Abdominal Pain dataset

| Diagnostic Group | With additional information Average accuracy | No additional information Average accuracy | Size of additional attributes |
|---|---|---|---|
| APP | 0.89±0.014 | 0.85±0.042 | 3 |
| DIV | 0.97±0.004 | 0.93±0.052 | 3 |
| PPU | 0.98±0.014 | 0.96±0.045 | 8 |
| CHO | 0.97±0.038 | 0.93±0.014 | 4 |
| INO | 0.95±0.016 | 0.91±0.009 | 3 |
| RCO | 0.94±0.022 | 0.93±0.016 | 6 |
| DYS | 0.89±0.080 | 0.86±0.029 | 2 |

**Table 3.** Predictions by SVM+ and SVM on Abdominal Pain dataset

| Diagnostic Group | SVM+ Average accuracy | SVM Average accuracy |
|---|---|---|
| APP | 0.88±0.009 | 0.86±0.021 |
| DIV | 0.63±0.015 | 0.60±0.019 |
| PPU | 0.54±0.010 | 0.53±0.013 |
| CHO | 0.82±0.005 | 0.79±0.028 |
| INO | 0.69±0.022 | 0.62±0.027 |
| RCO | 0.68±0.033 | 0.68±0.041 |
| DYS | 0.78±0.005 | 0.75±0.016 |

Experimental results are given in Table 2 where the binary classification is performed in one against all other classes. In batch learning mode, we only care about accuracies of predictions. To avoid the influence of redundant attributes, we use some selected symptoms here. For each disease group, we use 5 most relevant symptoms selected in [7] as "usual" attributes because these 5 selected symptoms could provide the similar confidence level as whole set of symptoms.

The features provided by experts in [1] are used as privileged attributes. The dataset is randomly divided into training set (4387 examples) and test set (2000 examples). The average accuracy and the corresponding standard deviation are shown for 7 diagnostic groups as the experts do not give any relevant information for the other two diagnostic groups (NAP and PAN). We then apply SVM and SVM+ on the same data, results are shown in Table 3. The kernel used here is Radial Basis Function(RBF), $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$, $\gamma \geq 0$. Cross-validation is applied on the training examples to find the optimal parameters. We can see that both SVM+ and our approach utilize additional information to improve classification accuracy. Due to the unbalance size of classes for prediction, accuracies of SVM and SVM+ are not as good as that of the conformal prediction approach.

## 5 Conclusion and Discussion

In this paper, we extend Conformal Predictors to deal with additional information. Experiments show that our approach successfully utilize additional information to improve the performance of classification as we expected. However, some more work need to be completed in the future.

We only used the Abdominal Pain dataset in this paper. Further experiments need be carried out on various databases. It would be interesting to consider and apply on-line predictions and slow learning where the feedback is given with an $n$-step delay. We would like to find out what kind of information could be defined as privileged.

## 6 Acknowledgements

This work was supported by EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots"); by EraSysBio+ grant funds from the European Union/BBSRC Shiprec project: "Living with uninvited guests"; by the National Natural Science Foundation of China (No.6112803) grant; and by grant "Development of New Venn Prediction Methods for Osteoporosis Risk Assessment" from the Cyprus Research Promotion Foundation. We would like to express our sincere thanks to Alex Gammerman, Vladimir Vovk and Vladimir Vapnik (Royal Holloway, University of London) for setting the problem, useful discussions and help.

## References

1. A. Gammerman and A.R. Thatcher. Bayesian Diagnostic Probabilities without Assuming Independence of Symptoms. Methods Inf Med. 30(1): 15–22, 1991.
2. A. Gammerman and V. Vovk. Hedging Predictions in Machine Learning. The Computer Journal, 50(2): 151–163, 2007.
3. H. Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. Tools in Artificial Intelligence: 315–329, 2008.

4. H. Papadopoulos, A. Gammerman and V. Vovk. Reliable Diagnosis of Acute Abdominal Pain with Conformal Prediction. Engineering Intelligent Systems 17(2-3): 127–137. CRL Publishing, 2009.
5. H. Papadopoulos, A. Gammerman and V. Vovk. Confidence Predictions for the Diagnosis of Acute Abdominal Pain. Artificial Intelligence Applications and Innovations III, IFIP International Federation for Information Processing, 296: 175–184. Springer, 2009.
6. F. Yang, H.-Z. Wang, H. Mi, C.-D. Lin and W.-W. Cai, Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics, 10(1): S22, 2009.
7. M. Yang, I. Nouretdinov, Z. Luo and A. Gammerman. Feature selection by conformal predictor. Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology, 364: 439–448, 2011.
8. L. Yu and H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning. Washington DC, 2003.
9. V. Vapnik and A. Vashist. A New Learning Paradigm: Learning Using Privileged Information. Neural Networks, 22: 544–557, 2009.
10. V. Vapnik, A. Vashist and N. Pavlovitch. Learning using hidden information: Master class learning. Proceedings of NATO workshop on mining massive data sets of security, 19: 3–14, IOS Press, 2008.
11. V. Vovk, A. Gammerman and G. Shafer. Algorithmic Learning in a Random World. Springer, 2005.