CONFORMAL PREDICTORS FOR RELIABLE PATTERN RECOGNITION

A. Gammerman
Royal Holloway, University of London
London, UNITED KINGDOM
e-mail: a.gammerman@rhul.ac.uk

Abstract

The talk reviews a modern machine learning technique called Conformal Predictors. The approach has been motivated by algorithmic notion of randomness and allows us to make reliable predictions with valid measures of confidence for individual examples. The developed technique guarantees that the overall accuracy can be controlled by a required confidence level. Unlike many conventional techniques the approach does not make any additional assumption about the data beyond the i.i.d. assumption: the examples are independent and identically distributed. The way to test this assumption is described. The talk also outlines some generalisations of Conformal Predictors and their applications to many different fields including medicine, cheminformatics, information security, environment, plasma physics, home security and others.

Keywords: conformal predictors, pattern recognition, data science

1 Background

The talk reviews a modern machine learning technique called Conformal Predictors [1, 2]. Given a set of training examples $(x_1, y_1), \dots (x_l, y_l)$, where each example consists of an object and a label, the problem of classification or regression can be considered as assigning a label y_{l+1} to a new object x_{l+1} , so that an example (x_{l+1}, y_{l+1}) does not look *strange* among the training examples. Or, in other words, how well the new example fits with the training set. In order to measure the *strangeness* of the new example in comparison with the training set, we introduced so-called *non-conformity measure* (NCM). This leads to a novel way to quantify the uncertainty of the prediction under rather general assumption. A non-conformity measure can, in principle, be extracted from any machine learning algorithm, such as SVM, logistic regression, neural networks, etc. We shall call the algorithm used for the extraction of an NCM as an underlying model.

Once an NCM is developed, it is possible to compute for any example (x, y) a $p - value$ that reflects how good the new example from the test set fits (or conforms with the i.i.d. assumption) with the training set. A more accurate and formal statement is this: chosen a significance level $\epsilon \in [0, 1]$ it is possible to compute $p - values$ for the test examples so that they are (in the long run) smaller or equal than ϵ with probability at most ϵ. Note that the key assumption here is that the examples in the training set and the test objects are independent and identically distributed (although a weaker requirement of exchangeability is sufficient). The idea is then to compute for a test object x a $p - value$ for every possible choice of the label y and make a prediction by choosing a label with the largest $p - value$ and the confidence as (1 - 2nd largest $p - value$). Once the $p - values$ are computed, they can be used in one of the following ways: a) to allow a user to specify a confidence level (or an error rate) so that the correct prediction rate is not worse than pre-specified confidence level; or b) to provide prediction with confidence for each individual example. More precisely:

- Given a significance level, ϵ, the predictor outputs *a region set* of possible labels for each test object such that the actual label appears no more than ϵ times in the set. This property is called *validity* of conformal predictors and it follows from the observations that in the online prediction protocol, the errors made $err_1^\epsilon, err_2^\epsilon, \ldots$ are independent and take value 1 with probability ϵ. Naturally, the narrower the prediction region is, the more *efficient* our prediction is

$$\Gamma^\epsilon = \{y \in Y : p(y) > \epsilon\},$$

  where Γ^ϵ is a prediction region, and the output provides the user with all labels y where $p - value$ is greater than ϵ.

- Another way is to supply a prediction for a new test object with two numbers: the **confidence**

$$\sup \{1 - \epsilon : |\Gamma^\epsilon| \leq 1\}$$

  and the **credibility**

$$\inf \{\epsilon : |\Gamma^\epsilon| = 0\}.$$

  Low credibility, for example, implies either the training set is non-random or the test object is not representative of the training set.

2 Conformal and Probabilistic Predictors

This method described above is so-called *transductive* conformal prediction (CP). It requires to retrain underlying model for each new test example. To make the method computationally more efficient, it has been generalised for *inductive* conformal predictor. In fact, there are now a number of various generalisations. Among them:

- *Inductive CP* (for computational efficiency). The inductive conformal predictors require the underlying model to be trained only once. The dataset is divided into *proper training* set, *calibration* set and *test* set. The proper training set is used only to calculate NCM scores (α's) of calibration and testing examples. Then $p - values$ are calculated using only those α's.

- *Mondrian CP* (for imbalanced data). In transductive and inductive CPs the examples we usually deal with belong to different classes or categories. Conformal predictors do not guarantee *validity* within the categories. The fraction of errors can be much larger than the pre-specified significance level for some categories, if this is compensated by a smaller fraction of errors in other categories. This validity within the categories is the main property of Mondrian conformal predictors[1]. Mondrian CP allows to have separate guarantees of the errors of different types. CP prediction set covers the true label with probability $1-\epsilon$. In Mondrian CP: if the true label is 1, then the prediction set contains 1 with probability $1-\epsilon_1$; if the true label is 0, then the prediction set contains 1 with probability $1-\epsilon_0$.

- *Probabilistic predictor* (produces reliable two-sided probabilistic estimates instead of p-values). Conformal predictors output $p-values$, but sometimes $p-values$ are more difficult to interpret than probabilities. In Bayesian decision theory: probabilities (but not the $p-values$) can be combined with utilities to arrive at optimal decisions. We have also developed a method of probabilistic prediction [1, 2] that is related to conformal prediction – so called Venn machine – that also has a guaranteed property of *validity.* It outputs multiprobabilistic predictions; for example, in the classification problem it provides a lower and upper bounds of probabilistic predictions.

Several other techniques have been developed such as *Cross-conformal* predictors (a hybrid of inductive CP and cross-validation); *On-line Compression Model* (for assumptions other than i.i.d.); *Conformal Predictive distribution* (provides the whole distribution and can be used for decision-making); *Ridge Regression Confidence Machine* and others.

The main point is that in all these generalisations the property of **validity** is preserved.

3 Applications

The conformal predictors techniques have been successfully applied in many fields: in medicine for diagnostic of ovarian and breast cancers; in neurosciences for diagnostic and treatment of depression; in information security in identifying various bots; in environment for assessing a level of pollution and many others. One of the most recent application is in pharmaceutical industry to find chemical compound activity using publicly available data [3]. A version of conformal predictors called Inductive Mondrian Predictor that keeps validity guarantees for each class has been applied for the large, high-dimensional, sparse and imbalanced pharmaceutical data. The experiments were conducted using several non-conformity measures extracted from underlying algorithms such as SVM, Nearest Neighbours and Naive Bayes. The results show that Inductive Conformal Mondrian Prediction framework allows to rank the compound activities and to find potentially useful molecules for drug developments.

[1] Called Mondrian because the categories resemble a Mondrian paintings by Piet Mondrian (1872-1944).

References

[1] Vovk V., Gammerman A., Shafer G. (2005). *Algorithmic learning in a random world*. Springer, New York.

[2] Gammerman A., Vovk V. (2007). Hedging prediction in machine learning. *The Computer J.* Vol. 50, pp. 151–163.

[3] Toccaceli P., Gammerman A. (2018). Combination of inductive mondrian conformal predictors. *Machine Learning*. Vol. 50, pp. 1–22.