Defensive Forecast for Conformal Bounded Regression

Ilia Nouretdinov, Alexander Lebedev

Defensive Forecast for Conformal Bounded Regression

Ilia Nouretdinov*, Alexander Lebedev**

Computer Learning Research Centre, Royal Holloway University of London*
Stavanger University Hospital, Centre for Age-Related medicine**

Abstract. The paper considers a conformal prediction method for bounded regression task. A predictor was based on the Defensive Forecast algorithm and has been applied for a medical prognostic problem. These empirical results are compared and discussed.

1 Introduction

The conformal prediction has been applied to regression estimation in [1–4], under assumption that label y is approximately linearly dependent on feature vector x. This was also extended for non-linear dependency using a non-linear transformation (kernel mapping) Φ of x into a higher dimensional space – see, for example, [9].

In this paper we consider the non-linear problem of bounded regression. A typical problem that requires a bounded regression is a prediction of examination marks, bounded from 0% to 100% or some problems in medical prognosis that have a range from healthy individuals to patients with a completely developed disease after a time delay. We apply an inductive conformal regression method to this type of problem to make valid regression estimations.

In Section 2 we recall key notions of machine learning and conformal prediction. In particular, what functions can be used as non-conformity measures and what changes if we apply conformal prediction in the inductive form of data processing that will be needed further. In Section 3 we describe a non-conformity measure for inductive conformal predictor based on K29 algorithm from [8], that was initially developed as game-theoretical approach (Defensive Forecast) to regression in bounded intervals. In Section 4 we give an example of application.

2 Machine learning background

2.1 Conformal classification and regression

The core element of a conformal predictor is a Non-Conformity Measure (NCM) that is a function A satisfying the equation

$$(\alpha_1,\ldots,\alpha_n) = A(z_1,\ldots,z_n) \Longrightarrow (\alpha_{\pi(1)},\ldots,\alpha_{\pi(n)}) = A(z_{\pi(1)},\ldots,z_{\pi(n)}).$$

NCM can be also undestood as a distance between a set $\{z_1,\ldots,z_n\}$ and one of its elements z_i, reflecting a relative strangeness α_i of the element with respect to the others.

The NCM values (non-conformity scores) are converted to p-values by the formula

$$p(z_1,\dots,z_n) = \frac{\#\left\{i = 1,\dots,n \mid \alpha_i^z \geq \alpha_n^z\right\}}{n}$$

A conformal predictor checks each of a set of hypotheses (possible labels) when presented with a new example and assigns it a p-value (Algorithm 1). Here $z_1,\dots,z_{n-1}$ are examples with known classification, each z_i consists of a feature vector $x_i \in X$ and the label $y_i \in Y$, and y is a hypothetical label for a new example with the feature vector x_n.

Algorithm 1 A step of conformal prediction

Input Non-conformity measure A
Input $z_1 = (x_1,y_1),\dots,z_{n-1} = (x_{n-1},y_{n-1}),x_{new}$
for $y \in Y$ **do**
  $z_n = (x_{new},y)$
  $(\alpha_1,\dots,\alpha_n) = A(z_1,\dots,z_n)$
  $p(y) = \frac{\#\{i=1,\dots,n|\alpha_i^z \geq \alpha_n^z\}}{n}$
end for

One of the ways to interpret p-values output by the conformal predictor is to find the prediction set R^γ is a list of labels that are not discarded at a given significance level γ:

$$R^\gamma = \{y : p(z_1,\dots,z_{n-1},(x_n,y)) > \gamma\}.$$

Conformal predictors are region predictors: their output is a prediction set R - a list of possible lanbales that are not discarded at a given significance level γ.

The prediction set should cover the true label y_n with probability at least $1-\gamma$, if i.i.d. assumption is true.

Alternatively the prediction can be done by comparing the different p-values and selecting more likely hypothesis. This makes results of conformal prediction comparable to standard ones if needed.

2.2 Inductive form of conformal prediction

The approach discussed above is a transductive version of conformal prediction. Inductive conformal predictor was proposed in order to make calculations more computationally efficient. Some previous applications of it can be found in [5, 6] and other works. Usually they use same non-conformity measures as standard (transductive) conformal predictors, but for this work we will need some extension of this.

The idea is to use a fixed additional set $u_1,\dots,u_h$ and to define the NCM $A(z_1,\dots,z_n) = (\alpha_1,\dots,\alpha_n)$ in such way that

$$\alpha_i = A_0(u_1,\dots,u_h,z_i).$$

Usually $u_1, \ldots, u_h$ (called *proper training set*) and $z_1, \ldots, z_{n-1}$ (called *calibration set*) are taken from the same data set with a random split. This interpretation of the inductive conformal framework is analogous to one given in [7] for inductive probabilistic (Venn) predictor.

A general scheme of Inductive Conformal Prediction (ICP) can be found in Algorithm 3.

3 Approach for bounded regression

Aim of this section is to present a conformal predictor based on K29 algorithm from the work [8] that develops a game-theoretic approach to machine learning. In principle, details related to this theory are not necessary to understand how the algorithms works and how non-conformity scores are calculated. However, we remind some of them in order to have some intuitive justification for the choice of non-conformity measure.

3.1 Prediction as a game

The Protocol 1 describes a simple form of prediction game with 3 players: Nature, Predictor and Sceptic. Nature generates examples x: say x_n on n-th round, Predictor gives a forecast $\hat{y}_n$ of Nature's move, and once he has done the prediction, Nature announces the real label y_n. Sceptic has an initial capital C_0 and bets s_n at round n.

Protocol 1
for $n = 1, 2, \ldots$ **do**
NATURE: x_n
PREDICTOR: $\hat{y}_n \in [-1, 1]$
SCEPTIC: s_n
NATURE: y_n
Sceptic's capital: $C_n = C_{n-1} + s_n(y_n - \hat{y}_n)$
end for

Usually in machine learning the Predictor tries to predict some value given by the nature (such a the new example's label) and his preformance is assessed by a loss function. However Nature does not have any interest to fail Predictor. Therefore game-theoretic approach to prediction usually assumes that Predictor has an antagonist, called Sceptic, whose win is Predictor's loss.

3.2 Defensive Forecast with non-conformity measure

In [8], Sceptic has to show in advance his potential reaction as a betting function S_n of Predictor's move, so that $s_n = S_n(\hat{y}_n)$.

Predictor develops the following strategy. After seeing the object x_n on round n Predictor has to solve the equation

$$\sum_{j=1}^{n-1} K((x_j, \hat{y}_j), (x_n, y))(y_j - \hat{y}_j) = 0$$

where K is a Mercer kernel.

Algorithm 2 follows K29 game protocol for the Defensive Forecast [8] and shows how the non-conformity scores $\alpha_n = |S_n(y_n)|$ could be extracted.

If Predictor's move (the prediction) is different from Nature's move (the label), then the discrepancy is measured not directly by their difference (as is it usually done in regression), but by the difference of Sceptic's reaction to them. This follows a general idea of game-theoretic probability: an event is rare or strange if someone with reasonable strategy may make a profit from betting for it. Therefore Sceptic's move showing his betting intention is used to measure the strangeness. If for example $S_n(y_n) = S_n(\hat{y}_n) = 0$, Sceptic prefers not to play in both cases, then the difference between y_n and $\hat{y}_n$ is not considered as an essential one.

Algorithm 2 K29 algorithm with players' strategies and NCM

Input: Kernel function $K((x,y),(x',y')) = \Phi(x,y) \cdot \Phi(x',y')$ where Φ is a continouous mapping to a Hilbert space.
for $n = 1,2,\ldots$ **do**
NATURE: x_n
SCEPTIC: $S_n(y) = \sum_{j=1}^{n-1} K((x_j,\hat{y}_j),(x_n,y))(y_j - \hat{y}_j)$
PREDICTOR: $\hat{y}_n \in [-1,1]$ is either y such that $S_n(y) = 0$ or the sign of S_n if it never reaches zero on $[-1,1]$.
NATURE: y_n
Sceptic's capital: $C_n = C_{n-1} + S_n(\hat{y}_n)(y_n - \hat{y}_n)$
NCM: $A_0((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),(x_n,y_n)) = |S_n(y_n)|$
end for

3.3 Using Defensive Forecast in inductive mode

NCM defined above can be used only in the *inductive* conformal prediction because otherwise, for transductive conformal predictors, the assumption of exchangeability does not hold: the order of the examples follows the protocol.

In the inductive mode of conformal prediction, the data are split into three parts of sizes h (proper training set $u_1,\ldots,u_h$), m (calibration set $z_1,\ldots,z_m$) and $N-h-m$ (testing set $z_{m+1},\ldots,z_{N-h}$). For an individual testing example the prediction is done as in Algorithm 3.

The only examples we deal with are the ones in the calibration or testing set, while proper training set can be considered as a parameter. That way the exchangeability property is satisfied.

If the non-conformity measure defined in Section 3.2 is applied in inductive mode, this means that Protocol 1 is run on examples $u_1,u_2,\ldots,u_h$ as usally, but the step $n = h+1$ is repeated many times starting from the same point. Each of calibration and testing examples in turn plays the role of x_{h+1} in Protocol 1 in order to get its non-conformity score. As for the testing examples, each of them is used also with different hypotheses about y_n, in this context a Nature's move on the step n may mean a hypothesis about this move.

Algorithm 3 A step of inductive conformal prediction

Input Non-conformity measure A_0
Input $u_1 = (x_{u.1}, y_{u.1}), \ldots, u_h = (x_{u.h}, y_{u.h})$
Input $z_1 = (x_1, y_1), \ldots, z_m = (x_m, y_m), x_{new}$
for $i = 1, \ldots, m$ **do**
  $\alpha_i = A_0(u_1, \ldots, u_h, z_i)$
end for
for $y \in Y$ **do**
  $z_{m+1} = (x_{new}, y)$
  $\alpha_{m+1} = A_0(u_1, \ldots, u_h, z_{m+1})$
  $p(y) = \frac{|\{i=1,\ldots,m+1 | \alpha_i^z \geq \alpha_{m+1}^z\}|}{m+1}$
end for

3.4 Kernels

In Algorithm 2 there is a parameter K: a kernel function (or a scalar product) after a *feature mapping* to a Hilbert space. This is analogous to the well-known kernels [9] $K(x,x') = \Phi(x) \cdot \Phi(x')$ but in K29 the kernels are dependent on y as well as on x.

This is useful for bounded regression problem because it allows to consider a non-linearity in a wider sense: non-linearity in y (labels) as well as labels rather than in x (feature vectors). An example is the polynomial kernel:

$$K_{Poly(d,e)}((x,y),(x',y')) = (x \cdot x' + 1)^d + (y \cdot y' + 1)^e.$$

where d and e are degrees of non-linearity in x and in y.

$S_n(y)$ can be represented as $\Phi(x_n, y) \cdot w_{n-1}$ where

$$w_{n-1} = \sum_{j=1}^{n-1} (y_j - \hat{y}_j) \Phi(x_j, \hat{y}_j)$$

plays a role similar to the slope w of separating hyperplane in Support Vector Machines [10]. But in SVM one can find w by solving a quadratic optimization problem, while in K29 calculation of w is separated into $n-1$ easy steps of on-line update.

Kernels depending on y were also used in a generalized form of SVM for structured output space [11] but in this algorithms optimization problem is even harder than in a standard SVM.

4 Application

4.1 Data

In our application, we use the data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, ADNI-1 cohort [17]. The database includes more than 800 subjects with up to 5 years annual follow-up with comprehensive clinical, neuropsychological, imaging and laboratory evaluations, performed at the specialized research centers. For the present study, we used 1.5 Tesla 3D T1 magnetic resonance

imaging (MRI) brain scans from patients with Alzheimer's Disease (AD), with Mild Cognitive Impairment (MCI) and Healthy Controls (HC), who had long term follow-up information and met the inclusion criteria (see *Diagnosis* below).

In earlier applications of conformal method to other MRI data (see [19]) the diagnostic was considered as a classification problem, while now we observe the data ordered by these labels reflecting the following disease stages.

- 164 healthy examples;
- 17 examples known to be healthy at the time of earliest measument who became Mild cognitive impairment (MCI) patients in less than 5 years time;
- 119 with Mild cognitive impairment (MCI);
- 156 known to be MCI at the measurement time and to convert to Dementia (AD) in less than 5 years time; this includes 62 examples will convert in at most 1 year;
- 169 with Dementia (AD).

4.2 Diagnosis

All AD patients met NINCDS/ADRDA criteria for probable AD, had mild level of dementia, defined as Mini-Mental State Examination (MMSE) score between 20 and 26, Clinical Dementia Rating Scale score of 1.0. Inclusion criteria for MCI were: 1) MMSE score between 24 and 30, 2) memory complaints and objective memory impairment measured by Logical Memory II subscale of the Wechsler Memory Scale (education adjusted), 3) CDR of 0.5, 4) absence of significant levels of impairment in other cognitive domains, 5) preserved activities of daily living, and 6) absence of dementia. MCI converters had to meet the criteria for Alzheimer's disease during at least two sequential evaluations (e.g., at 24 and 36 month follow ups). Controls (general inclusion/exclusion criteria): 1) MMSE scores between 28 and 30, 2) CDR of 0, 3) they did not meet criteria for clinical depression at baseline, MCI or dementia within 3 years of follow-up.

4.3 Image Post-Processing

Raw 3D T1 MRI data underwent `Freesurfer v5.1` (`http://surfer.nmr.mgh.harvard.edu`) steps for surface-based cortex reconstruction and volumetric segmentation. As a result, 68 measures of brain cortical thickness (32 for each hemisphere) averaged by parcellation as described in [15] and 41 volumetric measurements of subcortical structures (corrected for intracranial volume) acquired for every subject were combined with apoE-allele carrying information, basic clinical evaluations (MMSE and Word-recall) and demographics (age, gender, education). Each example therefore contained 109 brain morphometric measurements combined with 6 non-imaging features. Originally they were serial: same patient can have several measurements at different follow-up timepoints. For each patient, we will use its first (earliest) measurement. The label is based on the current diagnosis at that moment together with information about later dynamics of the disease.

4.4 Prediction intervals

According to the data structure, we consider the following 21 hypotheses related to ADNI.

- Healthy ($y = -1$);
- 4.5,4,...,1,0.5 years before Mild Cognitive Impairment (MCI) ($y = -0.9, \ldots, -0.1$);
- MCI non-converter ($y = 0$);
- MCI converter 4.5,4,...,1,0.5 years before conversion to Dementia ($y = 0.1, \ldots, 0.9$);
- Dementia ($y = +1$).

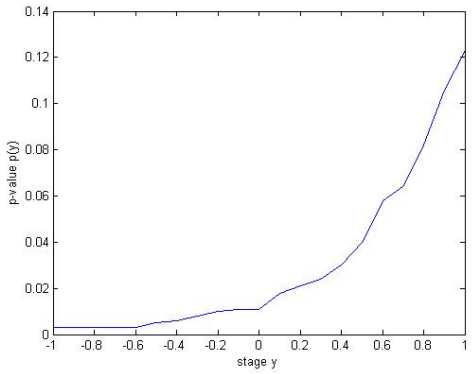

Fig. 1. Prediction for one of the examples: p-value as a function showing likelihood of a stage.

In order to apply K29 algorithm we use 109+6 features as vectors x and the stage numbers as their labels y. They are ranging from -1 to 1 with step 0.1 as shown in the list above.

Conformal predictor assigns p-value to each hypothesis about the diagnosis. A standard interpretation of conformal prediction is done in terms of intervals. Suppose that for one of examples, each possible y is assigned a p-value by the conformal predictor.

Fig. 1 presents a typical individual prediction made for an example. Its true label 0.9 meaning: MCI in 6 months before its conversion to AD.

Examples of corresponding prediction sets (intervals) are:

- for the significance level $\gamma = 10\%$, $R = \{y : p(y) > 0.1\} = [0.9; 1]$ that covers the true value with probability at least 90%;
- for the significance level $\gamma = 5\%$, $R = \{y : p(y) > 0.05\} = [0.6; 1]$ that covers the true value with probability at least 95%.
- for the significance level $\gamma = 1\%$, $R = \{y : p(y) > 0.1\} = [0.2; 1]$ that covers the true value with probability at least 99%;

4.5 Accuracy of two-class problems

In addition to prediction intervals, we can use p-values obtained form a conformal predictor for some two-class problems. The following ones were selected because of their popularity in the literature [13, 14, 16]:

- (A) Healthy vs Dementia;
- (B) Healthy vs MCI;
- (C) MCI non-converters vs (0.5–4.5 year) MCI converters
- (C_1) MCI non-converters vs (0.5-year and 1-year) MCI converters.

These problems can be solved by comparing highest p-values reached on corresponding intervals. For example, if we restrict our interest to the problem (C_1) then the interpretation of p-values is following:

- a prediction is correct in one of the following cases:

$$y = 0, p(0) > \max_{0.8 \le y < 1} p(y);$$

$$y \in \{0.8, 0.9\}, p(0) \le \max_{0.8 \le y < 1} p(y);$$

- wrong predictions:

$$y = 0, p(0) \le \max_{0.8 \le y < 1} p(y);$$

$$y \in \{0.8, 0.9\}, p(0) > \max_{0.8 \le y < 1} p(y);$$

- examples with true labels $y < 0; 0 < y < 0.8; y = 1$ are irrelevant for the accuracy although they are still used for training.

The best results are presented in Table 1.The accuracy is averaged over 50 random splits with ICP parameters $h = 500$ and $m = 100$ (see Sec.3.3). We also compare K29 with a simpler approach based on linear regression extended with a T-test feature selection step used in our previous work [19] applied in leave-one-out mode.

Underlying algorithm	Parameter	Task			
		(A)	(B)	(C)	(C_1)
Linear regression with feature selection		0.94 (best)	0.72 (best)	0.70	0.76
K29	trivial kernel	0.91	0.65	0.69	0.75
K29	polynomial kernel $K_{Poly(3,1)}$	0.50	0.42	0.57	0.32
K29	polynomial kernel $K_{Poly(1,3)}$	0.92	0.63	0.72 (best)	0.78 (best)

Table 1. Results with two-class accuracy

5 Discussion and Conclusions

This bounded conformal regression method has been applied to a problem of medical prognosis. A development of Alzheimer's disease has several stages before the actual dementia onset. Neurodegeneration usually starts from the entorhinal cortex and hippocampal formation and subsequently spreads thoughout the brain. This pattern is consistent with our results. Thus, the most important features for prediction were volumes of the Left and Right Hippocampi, Left Amygdala, thickness of the Left Entorhinal cortex, apoE-genotype (known genetic biomarker associated with different risks for Alzheimer's disease [18], and the result of Mini-Mental State Examination (screening tool to assess cognitive functions).

We have proposed a conformal predictor based on a new kind of non-conformity measure, based on the ideas of game-theoretic defensive forecasting method, originally developed for a bounded regression. This techniques has some advantages that were discussed in the theoretical part of the paper. The experimental results are especially interesting as an illustration of a generalized kernel technique in the context of bounded regression.

6 Acknowledgements

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 EraSysBio+ grant funds from the European Union/BBSRC Shiprec project: "Living with uninvited guests"; by Veterinary Laboratories Agency (VLA) of Department for Environment, Food and Rural Affairs (Defra) through the project: "Machine learning algorithms for analysis of large veterinary data sets"; 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 indebited to Alex Gammerman for setting up the problem and the idea of application. We would like to express our sincere thanks to Vladimir Vovk and Alexey Chervonenkis for useful discussions and help. Alexander Lebedev was supported by the Helse Vest Strategic Funding 2013. We would also like to thank Andrew Simmons (Institute of Psychiatry, King's College London) and Eric Westman (NVS, Karolinska Institute, Stockholm, Sweden) for their contributions to image collection and post-processing and Alzheimer's Disease Neuroimaging Initiative (ADNI) for making available neuroimaging data.

References

1. Vovk, V., Gammerman, A., Shafer, G. Algorithmic Learning in a Random World. Springer, 2005.
2. Melluish, T., Saunders, C., Nouretdinov, I., Vovk, V. Comparing the Bayes and typicalness frameworks. Lecture Notes in Computer Science. Vol. 2167, 2001, 360–371.
3. Vovk, V., Nouretdinov, I., Gammerman, A. On-line predictive linear regression. Annals of Statistics. 37(3), 2009, 1566–1590.

4. Gammerman, A., Vovk, V., Hedging Predictions in Machine Learning, Computer Journal, Vol. 50, Is. 2, 2007, pp. 151–172.
5. Papadopoulos, H., Haralambousm, H. Reliable Prediction Intervals with Regression Neural Networks. Neural Networks 24(8): 842–851. Elsevier, 2011.
6. Papadopoulos, H., Vovk, V., Gammerman, A. Regression Conformal Prediction with Nearest Neighbours. J. Artif. Intell. Res. (JAIR) 40: 815–840. 2011.
7. Lambrou, A., Papadopoulos, H., Nouretdinov, I., Gammerman, A. Reliable Probability Estimates Based on Support Vector Machines for Large Multiclass Datasets. Artificial Intelligence Applications and Innovations - AIAI International Workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB, Proceedings. II. Halkidiki, Greece, 2012, 182–191.
8. Vovk, V. On-line regression competitive with reproducing kernel Hilbert spaces. arXiv:cs/0511058v2
9. Kernel methods. Wiki for On-Line Prediction. `http://onlineprediction.net/?n=Main.KernelMethods`
10. Cristianini, Nello, Shawe-Taylor, John. An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.
11. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun., Y. Support Vector Learning for Interdependent and Structured Output Spaces. ICML, 2004.
12. Alzheimer's Disease Neuroimaging Initiative. Sharing Alzheimer's Research Data with the World. `http://adni.loni.ucla.edu/`
13. Liu, M., Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging Initiative. Ensemble sparce classification of Alzheimer's disease. Neuroimage, 60, 2012, 1106–1116.
14. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Alzheimer's Disease Neuroimaging Initiative. Multimodal Classification of Alzheimer's Disease and Mild Congitive Impairment. Neuroimage, 55(3), 2011, 856–867.
15. Destriux, C., Fischl, B., Dale, A., Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage, 53, 2010, 1–15.
16. Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernard, A., Narayan, V.A. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology 2012, 12:46 doi:10.1186/1471-2377-12-46.
17. Aisen, P.S., Petersen, R.C., Donohue, M.C., Gamst, A., Raman, R., Thomas, R.G., Walter, S., Trojanowski, J.Q., Shaw, L.M., Beckett, L.A., Jack, C.R., Jagust, W., Toga, A.W., Saykin, A.J., Morris, J.C., Green, R.C., Weiner, M.W., Alzheimer's Disease Neuroimaging Initiative. Clinical Core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. Alzheimers Dement., 2010, 239–246.
18. Alonso Vilatela, M.E., Lopez-Lopez, M., Yescas-Gomez, P. Genetics of Alzheimer's disease. Arch Med Res, 43, 2012, 622–631.
19. Nouretdinov, I., Costafreda, S.G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., Fu, C.H.Y. Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage, 56(2), 2011, 809–13.