DOBRO: A Prediction Error Correcting Robot Under Drifts

Alexandr Maslov *
TU Eindhoven
a.maslov@tue.nl

Hoang Thanh Lam
IBM Research
t.l.hoang@ie.ibm.com

Mykola Pechenizkiy
TU Eindhoven
m.pechenizkiy@tue.nl

Eric Bouillet
IBM Research
bouillet@ie.ibm.com

Tommi Kärkkäinen
University of Jyväskylä
tommi.karkkainen@jyu.fi

ABSTRACT

We propose DOBRO, a light online learning module, which is equipped with a smart correction policy helping making decision to correct or not the given prediction depending on how likely the correction will lead to a better prediction performance. DOBRO is a standalone module requiring nothing more than a time series of prediction errors and it is flexible to be integrated into any black-box model to improve its performance under drifts. We performed evaluation in a real-world application with bus arrival time prediction problem. The obtained results show that DOBRO improved prediction performance significantly meanwhile it did not hurt the accuracy when drift does not happen.

CCS Concepts

•**Information systems** → *Data mining;*

Keywords

Concept drift; On-line prediction error correction; ARIMA

1. INTRODUCTION

Our work focuses on correcting prediction errors made by a black-box model. As a motivational example let us consider a real-world bus arrival time prediction application currently deployed in the city of Dublin. The application gets real-time updates about the GPS locations of the buses and makes predictions of the buses arrival time at the next bus stops.

The current method for prediction is the kernel regression (KR) [4], which matches the currently observed trajectory with historical trajectories to predict the bus arrival time. The KR method is simple and effective, i.e. it was shown to work well for this problem. However, since it makes prediction by looking at historical data, it does not capture the non-recurrent incidents, e.g. accidents or unplanned events, happening in the current run of the bus. Under such drift circumstance, the prediction becomes inaccurate.

In general, the prediction performance of the black-box model is highly biased to the set of data collected at the moment of doing data analysis. When the built prediction model is deployed to run on the most recent data, it can not incorporate the changes that happen due to the modification in physical processes that generate the data. The gap between modelling and deployment happens quite often in practice, for example, modelling of the electricity load may become inaccurate when a new power plan is added to the grid, or modelling of transportation built on one city's data does not work well in another city with much different road network.

In this work, we propose a drift detection robot called DOBRO which continuously monitors the prediction residuals and performs correction when it is needed. We propose an effective policy which selects the relevant model for prediction correction and decide whether correction should be done depending on how likely the correction will lead to a better prediction. Thus we try to ensure that the correction will result in a better prediction performance and does not hurt the original prediction when the black-box model already does a good job.

DOBRO and related work. Our approach is based on the model output correction idea, i.e. the model output can be adjusted based on some addition information. An adaptive context-aware model-output correction for driver's route recognition has been proposed in [3]. Meta-learning that characterizes the performance of predictors on a problem at hand can be used for adaptive selection of predictors [6]. We also exploit the idea that the residuals of time series models are not independent and are predictable to the level that the model output correction becomes feasible. In concept drift research (see e.g. [2] for a recent survey) an idea of drift predictability in peer-to-peer settings has been explored [1].

The paper is organized as follows. In Section 2 we formally define the problem. In Section 3 we present our method. We demonstrate the effectiveness of the method in experiments in Section 4. Section 5 concludes the paper.

*A. Maslov is affiliated with both TU Eindhoven and University of Jyväskylä within the double PhD agreement. This work was conducted during his visit to IBM Research Dublin.

SAC 2016,April 04-08, 2016, Pisa, Italy

http://dx.doi.org/10.1145/2851613.2851888

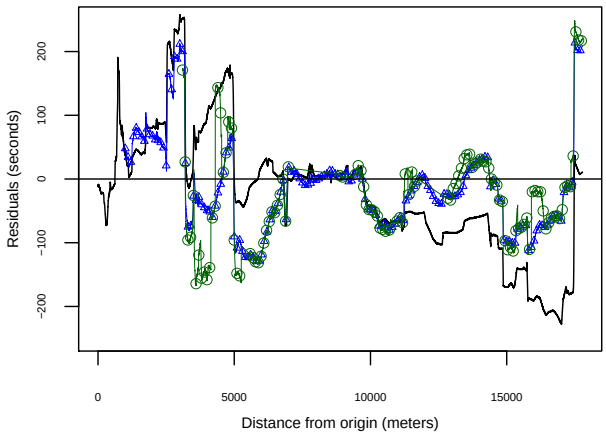

Figure 1: Prediction residuals of a bus arrival time using the KR method with the prediction horizon 1000 meters ahead. Blue triangles and green circles denote corrected residuals.

2. PROBLEM FORMULATION

Let $\langle y_t \rangle \equiv \langle y_1, \ldots, y_T \rangle$ be the series of observed at time moments $t \in \langle 1, \ldots, T \rangle$ values of the response variable $\mathbf{y}$, e.g. bus arrival time. Denote $\langle \hat{y}_t \rangle$ as a series of black-box model's predictions made with prediction horizon h. Denote $\langle r_t \rangle$ as the residuals, i.e. $r_t = \hat{y}_t - y_t$.

A good model always results in the residuals that resemble a random noise series with mean value equal to zero. However, the distribution of residuals may deviate far from random noise due to various reasons. For instance, context has changed when the physical process that generates data has changed. Under such circumstance residual series may have strong patterns such as autocorrelation enabling us to predict the residuals and make correction.

Assume that at time moment t the black-box model gives prediction of the response variable as $\hat{y}_{t+h}$. The corresponding residual of prediction equals to $r_{t+h} = \hat{y}_{t+h} - y_{t+h}$. Assume that we have a method that predicts the residual as $\hat{r}_{t+h}$. The corrected prediction is defined as:

$$y^*_{t+h} = \hat{y}_{t+h} - \hat{r}_{t+h}.$$

If prediction of the residual works well i.e. $\hat{r}_{t+h}$ is very close to r_{t+h} the correction y^*_{t+h} will be a better prediction of y_{t+h}. However, a fundamental question is how close the value $\hat{r}_{t+h}$ to the value r_{t+h} should be so that the correction gives benefit? The following principle called as correction principle shows the necessary condition that ensures positive benefit when correction is performed:

CORRECTION PRINCIPLE: The necessary condition that ensures reduction in prediction error after correcting the prediction with an estimate of residual is:

$$|r_{t+h} - \hat{r}_{t+h}| < |r_{t+h}|. \quad (1)$$

It is trivial to prove the correction principle because ineq. 1 is equivalent to:

$$|y^*_{t+h} - y_{t+h}| < |\hat{y}_{t+h} - y_{t+h}|.$$

It is important to note that the value $|r_{t+h} - \hat{r}_{t+h}|$ is equal to the absolute value of the residual of the corrected prediction. Denote δ_{t+h} as the difference between the residuals before and after correction, i.e.

$$\delta_{t+h} = |r_{t+h}| - |r_{t+h} - \hat{r}_{t+h}|. \quad (2)$$

The larger the value of δ_{t+h} is the more benefit we get from prediction correction.

An interesting result of the correction principle is that prediction of the absolute value of the residual does not need to be perfectly accurate to ensure the benefit of making correction if we capable to predict sign of residual with a high accuracy. E.g. assume that the residual is equal to 20; if the predicted residual is between 0 and 40 then making correction is still beneficial. In particular if the predicted residual is 30 then the benefit of making correction is equal to $20 - |30 - 20| = 10$. This simple example encourages us to perform prediction correction when we expect that ineq. 1 holds.

The problem can be approached as follows. We need to develop (1) a method to estimate the residual $\hat{r}_{t+h}$ at time point t; and (2) a correction policy that can estimate the probability that ineq. 1 is valid and decide whether we should perform correction.

3. APPROACHES

3.1 Residual prediction models

To predict residual values $\hat{r}_{t+h}$ we used several popular forecasting methods: Random walk (Naive), Moving average (MA) and Linear trend model (LM). All three models are special cases of the ARIMA model which in general case can be written as

$$\phi(B)\nabla^d r_t = \theta(B)a_t + c, \quad (3)$$

where $B^m r_t = r_{t-m}$ is a backward shift operator, $\nabla^d \equiv (1-B)^d$ is a backward difference operator of order d, a_t is an error term, and c is a constant value. In case of a stationary process, r_t is replaced by $\widetilde{r_t} = r_t - \mu_r$ where μ_r is a mean value. Polynomials $\phi(B)$ and $\theta(B)$ define autoregressive (AR) and moving average (MA) models.

Naive model corresponds to the ARIMA(0,1,0) process, called random walk without drift model. AR and MA terms are:

$$\phi(B) = 1,\ \nabla^d = (1-B),\ \theta(B) = 1,\ c = 0$$

Moving average model is equivalent to the ARIMA(0,0,0) stationary process when r_t can be replaced by difference with a mean value $\widetilde{r_t} = r_t - \mu_{r_t} = a_t$. AR and MA terms in that case are:

$$\phi(B) = 1,\ \nabla^d = 1,\ \theta(B) = 1,\ c = 0$$

LM model corresponds to the random walk model with the drift, or equivalently to ARIMA(0,1,0) process with the constant non-zero term $c \neq 0$.

Our preliminary experiments show that complex ARIMA models with a large number of parameters easily lead to overfitting and predictions obtained using these models are highly unstable due to high variability and non-stationarity of the data within sliding window. Also we observed that Linear model often under- or over-predicts residuals. Since

correction obtained using linear regression and ARIMA models of a high order worsens predictions made by the black-box model we excluded these models from our framework.

3.2 Continuous correction policy

Continuous correction strategy corrects black-box predictions at every time step using correction terms $\hat{r}_{t+h}$ estimated using the models described by the generic Equation 3. To select the best model we use selection procedure consisting of three steps: (1) selecting the model with the best performance (BM) during recent time interval, (2) testing if the gain obtained by selected model is positive, (3) if the gain was positive we make correction on the next step using BM, otherwise we don't make correction. The best model is the model with the highest gain value during the past k moments of time: $g[i] = |r_i| - |r_i - \hat{r}_i|$, $i = (t-k, \ldots, t)$. To decide whether to do correction or not we performed t-test with $H_0 : \mu_g = 0$ and $H_1 : \mu_g > 0$. If p-value $p > 0.05$ H_0 is rejected and we make prediction with the selected model.

3.3 Smart correction policy

Under the smart correction policy, the decision to make correction is made on the basis of estimated probability of beneficial correction. To predict residuals we make two assumptions: (1) it makes sense to try to correct black-box predictions by estimated residual values if we observe pattern in their behavior in a form of relatively long drifts below/above zero level, (2) the distribution of residuals may change and their behavior can be described by a Gaussian random walk model, which is a particular type of Markov stochastic process. This is a simple yet generic model which is widely used in many scientific applications. The model implies two properties: (1) the values of residual changes $\Delta r = r_t - r_{t-1}$ are independent for any two different time moments t, (2) the change $\Delta r_t = \epsilon_t$ during a discrete time interval Δt has a normal distribution $\mathcal{N}(0, \sigma)$ [1] Thus we estimate probability of making beneficial correction at the future moment $t + h$ using the current residual estimate $\hat{r}_{t+h}$.

The change in the residual value after period of time h is a sum of changes during a smaller discrete time intervals constituting h

$$r_{t+h} - r_t = \sum_{i=1}^{h} \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma) \tag{4}$$

The sum of two independent normally distributed variables is also normally distributed variable with the mean and variance equal to the sum of two means and variances. Thus probability distribution of the change in residual value after h moments of time is:

$$P(r_{t+h} - r_t) = \mathcal{N}(0,\ \sigma\sqrt{h}) \tag{5}$$

Parameter σ can be estimated by calculating standard deviation of the residual changes at successive moments of time in historical data.

From the correction principle it follows that the correction is beneficial if $r_{t+h} > \frac{\hat{r}_{t+h}}{2}$ when $\hat{r}_{t+h} > 0$, and if $r_{t+h} < -\frac{\hat{r}_{t+h}}{2}$ when $\hat{r}_{t+h} < 0$. Therefore, the probability of beneficial correction $Pr(\hat{r}_{t+h})$ for current residual estimate $\hat{r}_{t+h}$ is given by:

$$\begin{aligned} Pr(\hat{r}_{t+h} > 0) &= 1 - \Phi(\frac{\hat{r}_{t+h}}{2}) \\ Pr(\hat{r}_{t+h} < 0) &= \Phi(\frac{\hat{r}_{t+h}}{2}) \end{aligned} \tag{6}$$

where $\Phi(r_{t+h})$ is estimated from the historical data, i.e. a cumulative distribution function defined by integral of probability distribution 5 with mean value r_t

$$\Phi(y) = \int_{-\infty}^{y} \mathcal{N}(y \mid r_t,\ \sigma\sqrt{h} \cdot k_\sigma) dy \tag{7}$$

Here we introduced an additional scale parameter k_σ which defines an uncertainty in estimation of σ. Equation 6 defines smart correction policy[2]. We make corrections only when the probability of a successful correction is higher than the predefined threshold s, i.e. $Pr(\hat{r}_{t+h}) > s$.

4. EXPERIMENTS WITH THE BUS DATA SET

Bus data set [3] is created from 1570 bus 18 kilometers long trajectories among which 90% of trajectories were used as a training set for KR and the rest as a test set. The KR method makes predictions for each meter on the trajectories [5].

In the first part of experiment we assess the continuous correction policy. For each data set we performed prediction-correction with three different prediction horizon values H and three sliding windows sizes W. To select the best model we calculated gain values for the past 100 meters, i.e. $k = 100$. The results are summarized in Table 1. The first value in each cell is a relative reduction in RMS, the second value is the precision[4] (both in %). In the vast majority of cases our approach did apply one of the corrections and only occasionally left the model output as is. The performance deteriorates with larger H. The results are not very sensitive to the choice of W.

Table 1: Performance of the continuous correction.

MA	W = 500	17.68 (66)	1.72 (60)	-16.71 (53)
	W = 1000	15.88 (64)	0.84 (58)	-15.73 (52)
	W = 2000	12.87 (60)	2.22 (57)	-15.05 (51)
Naive	W = 1	29.73 (73)	10.86 (65)	-11.88 (56)
BM	W = 500	20.96 (70)	5.57 (61)	2.97 (54)
	W = 1000	21.50 (70)	6.79 (61)	2.47 (54)
	W = 2000	21.92 (71)	10.88 (61)	-0.40 (53)

In the second part of the experiment we assess the smart correction policy. We used Naive model as a correction model since it is the best in terms of number of successful corrections (Table 1).

[1] Further we use notation $\mathcal{N}(\mu, \sigma)$ for random variable and $\mathcal{N}(x|\mu, \sigma)$ for probability density function.

[2] Dobro source code is available in Github repository github.com/av-maslov/Dobro

[3] Samples of the Dublin Bus GPS data are available on the web site:dublinked.com/datastore/datasets/dataset-291.php

[4] Proportion of performed corrections which are successful.

If the smart corrector makes not beneficial correction (i.e. ineq. 1 is not satisfied) we count this event as a False Positive (FP). If correction is beneficial we count it as a True Positive (TP). If the smart corrector decides not to do correction but it turn's out that correction made using continuous corrector is beneficial we count this event as False Negative (FN). In the opposite situation - as a True Negative (TN).

Results of the experiment are shown on the plot A in Figure 2 in a form specificity-sensitivity plot obtained by varying coefficient k_σ. Performance of continuous corrector is a point (0,1) since number of TN and FN is zero. Solid and dashed lines depict averaged performance for prediction horizons $h = 500, 1000$ over 157 bus routes depicted by gray lines.

In most of the cases decisions to switch corrector on/off were beneficial. Performance decreases with larger values of prediction-correction horizon. Plot C in Figure 2 illustrates an output result of the corrector with a parameters corresponding to the point L on the plot A. Vertical white regions are moments when continuous corrector reduced accuracy. Black arrows starting at zero level depict moments when decision was to switch correction off.

Third part of experiment. After the decision to make correction is made the next step is to estimate the value by which to correct. Performance of this step can be measured by counting cases when RMS error is reduced (Win), increased (Loss) or left the same (Draw). As a baseline we calculated the same metrics for the continuous and random correctors. The latter one makes correction at random using uniform random generator.

Results for these performance metrics are shown on the plot B in Figure 2. White and filled black symbols denote wins and losses correspondingly. Results for continuous corrector are depicted by the square symbol, for random corrector - by circles and for smart corrector - by triangles. Performance of the continuous and random correctors does not depend on k_σ coefficient and therefore have constant values depicted by horizontal lines. Fraction of draws is not depicted for clarity but it can be easily calculated as 1 minus sum of fractions of wins and losses. Rectangle encloses cases when smart corrector has lower loss-rate than a continuous corrector and still higher win-rate than a random corrector.

It can be seen that continuous corrector is optimal in terms of wins and random corrector is optimal in terms of losses. Smart strategy is a trade-off between them. Smart policy will be optimal when the cost for wrong correction is higher than for reduced cost after successful correction.

5. CONCLUSIONS AND FUTURE WORK

We proposed a lightweight online prediction error correcting mechanism that can operate under drifts. Our experiments suggest that even with a simple continuous correction policy we can achieve considerable improvement in terms of reduced RMS error of (corrected) predictions made by the black-box models.

DOBRO correction policy switches continuous corrector on and off in a smart way according to the expected behavior of residuals. Our experimental study confirms that it is safe to correct predictions even from a good model which has residuals in a form of random noise because correction errors will compensate each other on average.

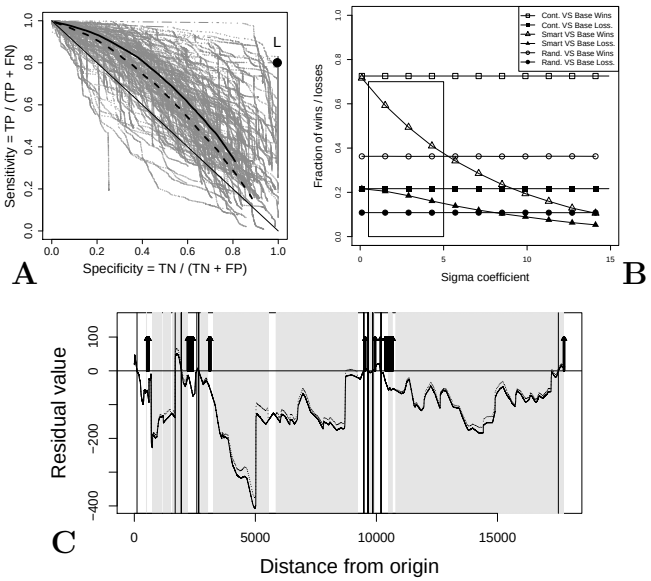

Figure 2: **A**: Average sensitivity-specificity. **B**: Comparison of continuous (squares), random (circles) and smart(triangle) error-correctors in terms of Wins(white symbols)/Losses(black symbols). **C**: An illustration of the prediction error correction using smart policy.

If residuals behavior is not random noise, but reveals some drifts above or below zero level, then DOBRO can be used to balance between more robust and more reactive prediction models.

We considered the settings in which residuals for past predictions become immediately available as input for our correction model. However, in practice there is often a delay in labeling. It is interesting to develop a correction policy that can deal with such delayed and/or irregular labeling.

Acknowledgments. This research is partly supported by STW CAPA project and COMAS.

6. REFERENCES

[1] H. H. Ang et al. Predictive handling of asynchronous concept drifts in distributed environments. *TKDE*, 25(10):2343–2355, 2013.

[2] J. Gama et al. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44, 2014.

[3] O. Mazhelis et al. Context-aware personal route recognition. In *DS'2011*, pages 221–235.

[4] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[5] M. Sinn et al. Predicting arrival times of buses using real-time GPS measurements. In *ITSC'2012*, pages 1227–1232.

[6] I. Zliobaite, J. Bakker, and M. Pechenizkiy. Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? *ESWA*, 39(1):806–815, 2012.