21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Temporal Feature Selection for Time-series Prediction

Shohei Hido* Tetsuro Morimura
IBM Research - Tokyo
hido.jp@gmail.com *tetsuro@jp.ibm.com*

Abstract

We present a feature selection method for multivariate time-series prediction. It aims to use the best sliding window size and delay for each explanatory variable, which are usually fixed. The idea is to convert the original time-series into a set of cumulative sum with different length. The combinations of cumulative sum variables obtaining nonzero weights in sparse learning algorithms represent the optimal temporal effects from explanatory variables to the target variable. Experiments show that the method performs better than conventional methods in regression problems.

1 Introduction

Time-series prediction includes a set of tasks to estimate the current state or category of time-varying objects based on the past measurements. Many applications in pattern recognition relate to time-series prediction due to its temporally-changing nature, such as gesture recognition on hand motion [4] and electrocardiogram classification for biometrics [7]. Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) have been widely-used for modeling time-series data taken from audio, vision, or human body which are usually univariate or low-dimensional. However, other applications such as driver's stress prediction from multiple sensors [5] are emerging, though HMM and DTW are both not good at handling such multivariate time-series prediction.

The common approach to this problem is based on stationary supervised learning algorithms with sliding window which regard fixed-length moving average for explanatory variables as i.i.d input variables. However, it cannot handle complicated temporal effects involving different time-windows, different times-lags, and decays between explanatory and target variables. Healey and Picard also used a stationary feature selection method that does not take into account temporal effects between the variables [5].

In this paper we propose a new method to choose the optimal time-windows and time lags for each variable based on feature pre-processing and sparse learning. The key is to convert the original time-series values into a series of cumulative values for different durations. Sparse learning algorithms outputs a model of which coefficients effectively represent the underlying temporal effects from each explanatory variable to the target variable.

2 Preliminary

If we regard each set of the values for both target and explanatory variables at a time step t as a data sample $\mathrm{z}_t = \{\mathrm{x}_t, y_t\}$, time-series prediction also belongs to supervised learning problem, though it is different from the ordinary classification or regression problems. The efficient way to make use of the past data samples is the main issue in time-series prediction. One of the most successful and widely-accepted approach is to use the moving average [2]. Figure1 shows an example of time-series regression problem. There are two explanatory variables, x_1 and x_2. y shows the time-series of the target variable. In order to taking into account the past values, first we convert them into the time-series of the average values within a fixed window size w, $\bar{x}_1$ and $\bar{x}_2$. However, real-world applications may include the underlying true model having more complex temporal effects of different time lags and window sizes for all variables as shown in Figure 2. A naive approach is to use all of the past values as independent explanatory variables as studied in the distributed lag problem in statistics [1]. However, to the best of our knowledge, there is no efficient way to choose the coefficients in case of high-dimensional multivariate time-series regression. Another approach is to compute the moving average values for all combinations of possible time lags and window sizes for each original explana-

[1]Currently with Preferred Infrastructure, Inc.

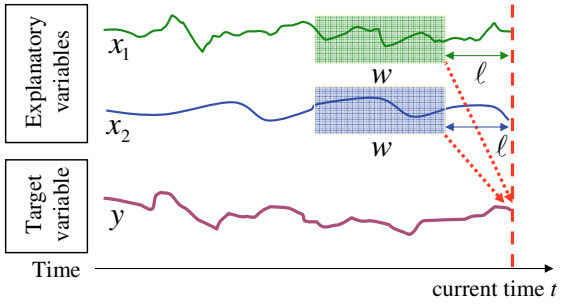


Figure 1. Time-series regression with fixed-length moving average.

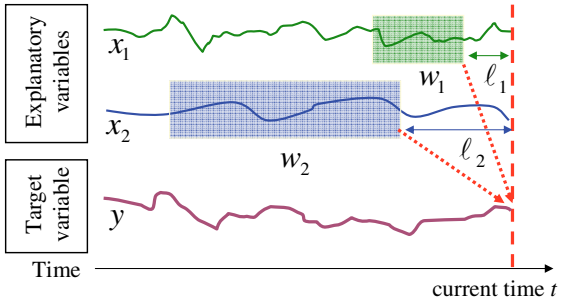


Figure 2. Time-series regression with different time-lags and time-windows.

tory variables in a brute-force manner. Though we can choose the best coefficients by using sparse learning, since the number of input variables becomes very large, it may lead to the model unstability or over-fitting problem.

3 Method

Figure 3 shows the two key steps in our method.

(1) Cumulative sum variable transformation. Given time-series values of an original variable, we compute cumulative sum variables for each time step t. We set the maximum possible time lag as L and the maximum possible window size as W. The cumulative sums are computed by increasing the gap g, which corresponds to the sum of the lag and window size, up to the maximum gap $(L+W)$. Then we obtain a set of the sums of size $L+W$ for each time step t as follows.

$$x_g^c(t) = x(t-1) + \ldots + x(t-g).$$

The total number of input variables as the cumulative sum variables equals $D(L+W)$ where the number of the original variables is D.

(2) L1-regularized sparse learning. Then we apply sparse learning algorithms for ordinary regression/classification problems on the cumulative sum variables of D variables. This allows us to effectively select the important features i.e. important temporal effects on the cumulative sum variables. In Figure 3(2), we assume that there are only two non-zero coefficients shown as red bars, for $x_\ell^c(t)$ and $x_{\ell+w}^c(t)$, with the op-

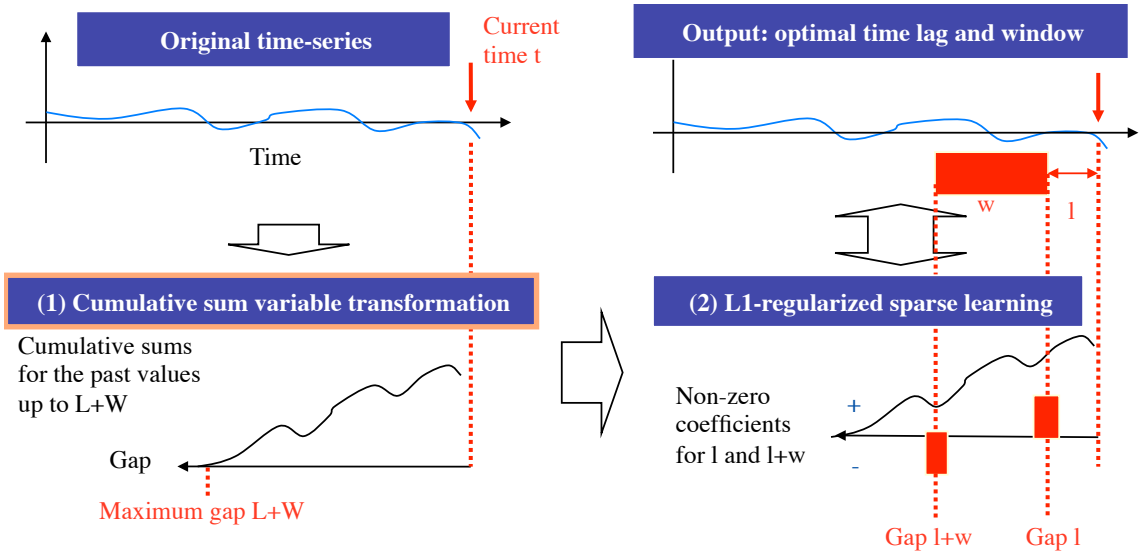


Figure 3. Two important steps in the proposed approach

posite signs and the same weight. The difference of these cumulative sum variables can represent the optimal time lag and window since the effects from the past values $x(t-1)$ to $x(t-\ell)$ are canceled and there remain only the sum of $x(t-\ell-1)$ to $x(t-\ell-w)$ as

$$\beta x^c_\ell(t) - \beta x^c_{\ell+w}(t) = -\beta(x(t-\ell-1)+\ldots+x(t-\ell-w)),$$

where β is the absolute value of the coefficient. This means that the optimal lag ℓ and the optimal window size w were derived at the same time. Note that this method can also work well to obtain a simple stepwise model for more complex cases with more than two non-zero coefficients.

Advantage of the proposed approach. The main advantage is the computational efficiency by using only $L+W$ variables for D original variables even though the degree of freedom is equivalent to that of the brute-force approach. Note that the true model in real-world applications would be more complex by involving increases and decays in the temporal effect represented by more non-zero coefficients. However, our method still works well in such cases by approximating the effects by a step-wise function which is represented by a few additional non-zero coefficients on the cumulative sum variables. In addition, this approach can also be applied other than moving average to second or higher-order moments (as a replacement of moving variance), and exponentially-weighted sums, though the application is beyond the scope of this paper.

4 Numerical results

We compare our method with naive and brute force approaches in regression problems. We assume that the training data set , the dimension D, the maximum possible lag L and window size W are given. In our method, we used only the cumulative sum variables of size $D(L+W)$. Naive approach involves only the delayed variables of size $D(L+W)$ as same as the distributed lag problem [1]. On the other hand, in the brute-force approach, we use the moving average values for all combinations of the candidate lags and windows (total DLW variables). For all of the methods, we used LARS (Least-angle regression) [3] as the sparse learning algorithm. The L1 regularization parameter was selected as to minimize Cp statistics [6].

4.1 Artificial dataset

We generate two explanatory variables of sin and cosine waves of different frequencies as $x_a = \sin(2t) + \epsilon$ and $x_b = \cos(t) + \epsilon$ where the error term $\epsilon \sim N(0, 0.5^2)$. The true regression model for target variable y is defined as,

$$y(t) = 1.3 * \mathrm{sw}(x_a, 5, 2) - 0.7 * \mathrm{sw}(x_b, 2, 8) + \epsilon,$$

where the temporal effect function $\mathrm{sw}(x, \ell, w)$ represents the value of a moving average with lag ℓ and window w. The task is to estimate the true model from the candidate set of lags and windows, $\ell = \{0, 1, \ldots, 5\}$ and $w = \{1, 2, \ldots, 10\}$.

Figure 4 and 5 show that the coefficients estimated by our methods. They are sparse and very close to those of the true models, which are also represented as coefficients on the cumulative sum variables. On the other hand, the brute-force method generates an unstable over-fitting model with many non-zero and larger coefficients as Figure 6 shows those on x_a. This leads to the difference in prediction accuracy. Figure 7 shows the root mean squared error (RMSE) for both approaches when changing the size of the training set from 50 to 500. The proposed method had much smaller error and avoided over-fitting caused by the larger size of training set, unlike the brute-force approach. In Figure 8, we show the running time required for both meth-

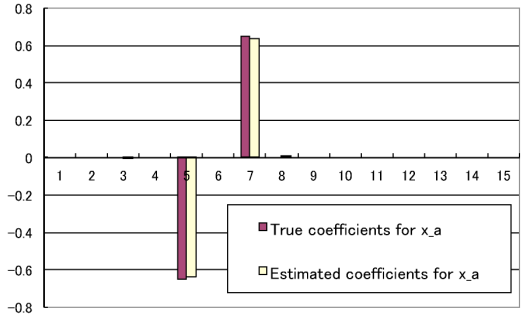


Figure 4. Estimated x_a coefficients

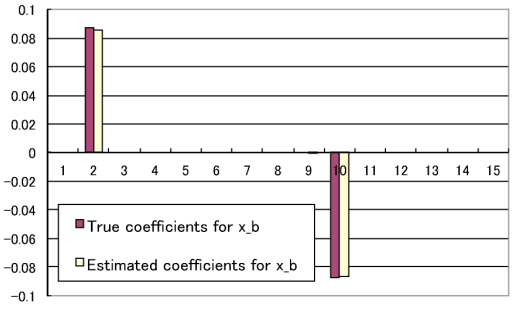
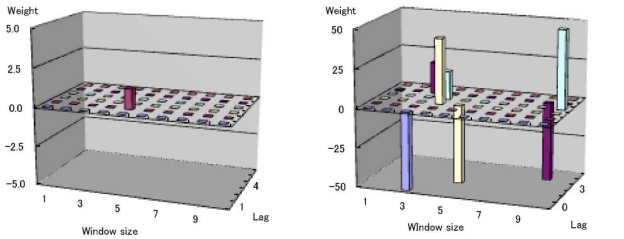


Figure 5. Estimated x_b coefficients



True coefficients for x_a Estimated coefficients for x_a

Figure 6. Estimated x_a coefficients on brute-force approach

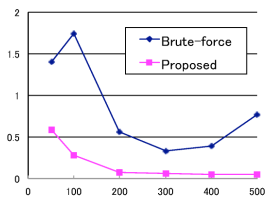


Figure 7. RMSE

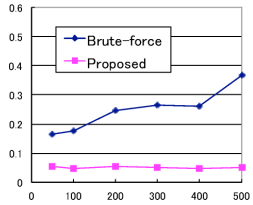


Figure 8. Time

ods. These indicate that the proposed method is also beneficial in terms of computational efficiency since the LARS can perform faster and more stable on the smaller and sufficient set of cumulative sum variables.

4.2 Physiological dataset

We also tested our method using a real-world physiological dataset [8]. The dataset includes three variables for a sleeping patient, the height of breast, the intensity of oxygen in blood, and the heat rate. We tried to predict the heart rate based on the past measurements of the remaining two variables and their standard deviation computed for last five samples. Therefore, the task is a time-series regression to estimate the heart rate with four explanatory variables. We used the first $16,900$ samples as the training set and the last 100 samples as the evaluation set. The same setting and lag/window candidates are the same with the artificial data experiment for all of the three approaches.

The RMSE and running time are shown in Table 1. Though the naive approach is faster, our method achieved smaller error. The brute-force approach showed comparable accuracy with our method, it performed much slower. In summary, our method achieved the best prediction performance by efficiently capturing the hidden dynamics between the variables with a reasonable computational cost.

Table 1. Results on physiological data

Approach	RMSE	Run. time (sec)
Naive	0.09243	0.02
Brute-force	0.08966	7.16
Proposed	0.08853	0.86

5 Conclusion

In this paper we addressed the problem of diverted sliding window size and delay on time-series prediction. We proposed the cumulative value transformation method for efficiently representing time-series data and introduced an algorithm based on sparse learning. We demonstrated that our approach outperforms existing methods in practice through experiments using real-world datasets.

Future work includes real-world applications of the method to time-series classification problems and further extensions for enabling higher-order features such as moving variance.

References

[1] S. M. Almon. The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196, 1965.

[2] G. E. P. Box and G. M. Jenkins. Time series analysis: forecasting and control. 1976.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[4] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis. A hidden markov model-based continuous gesture recognition system for hand motion trajectory. In *Proc. of ICPR'08*, pages 1–4, 2008.

[5] J. Healey and R. Picard. Smartcar: Detecting driver stress. In *Proc. of ICPR'00*, pages 218–221, 2000.

[6] C. L. Mallows. Some Comments on Cp. *Technometrics*, 15(3):661–675, 1973.

[7] V. N and S. Jayaraman. Human electrocardiogram for biometrics using dtw and flda. In *Proc. of ICPR'10*, pages 3838–3841, 2010.

[8] A. S. Weigend and N. A. Gershenfeld. *Time series prediction: Forecasting the future and understanding the past*. Addison-Wesley, 1994.