Information Sciences 375 (2017) 219–233

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Markov cross-validation for time series model evaluations

Gaoxia Jiang, Wenjian Wang*

School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China

ARTICLE INFO

Article history:
Received 21 July 2016
Revised 5 September 2016
Accepted 28 September 2016
Available online 29 September 2016

Keywords:
Model evaluation
Markov cross-validation
Time series

ABSTRACT

Cross-validation (CV) is a simple and universal tool to estimate generalization ability, however, existing CVs do not work well for periodicity, overlapping or correlation of series. The corresponding three criteria aimed at describing these properties are presented. Based on them, we put forward a novel Markov cross-validation (M-CV), whose data partition can be seen as a Markov process. The partition ensures that samples in each subset are neither too close nor too far. In doing so, overfitting model or information loss of series, which may result in underestimation or overestimation of the error, can be avoided. Furthermore, subsets from M-CV partition could well represent the original series, and it may be extended to time series or stream data sampling. Theoretical analysis shows that M-CV is the unique one which meets all of above criteria among current CVs. In addition, the error estimation on subsets is proved to have less variance than that on original series, therefore it ensures the stability of M-CV. Experimental results demonstrate that the proposed M-CV has lower bias, variance and time consumption than other CVs.

1. Introduction

Time series appears in many fields, e.g. economics, meteorology, finance, medicine and many others. Time series processing techniques mainly include prediction, smoothing, regression and others. Time series prediction (auto-regression) aims to forecast future values by the past series. Time series smoothing is to build an approximating function that attempts to capture important patterns of series. And time series regression is to create a functional relation between the response series and exogenous variables. Evaluating the performance of models of time series is an important problem when choosing the better one among various available models or parameters. Many aspects of models, e.g. generalization ability or error, complexity, interpretability, should be considered. For time series models, generalization error may be the most important factor, so most literatures about comparing time series models focus on it. The key problem for comparison of generalization ability is how to estimate generalization error.

There are some traditional approaches to estimate generalization error at present. Hold-out is an estimator with low computational complexity. Its downside is that the results are highly dependent on the choice for data split [20]. The bootstrap estimator is known to have better performance on small samples. However, in all situations of severe overfit, the estimator is downwardly biased [6]. Cross-validation is an estimator widely used to estimate generalization error for its practicability and flexibility. The above estimators have been compared in related researches [10,11]. Kohavi [11] studied above methods, and the results indicated that the best method for model selection is 10-fold stratified cross-validation. Kim [10] performed an empirical study to compare the 0.632C bootstrap estimator with the repeated 10-fold cross-validation

* Corresponding author.
E-mail addresses: jianggaoxia@sxu.edu.cn (G. Jiang), wjwang@sxu.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.ins.2016.09.061
0020-0255/© 2016 Elsevier Inc. All rights reserved.

and the repeated one-third hold-out estimator, and the results showed that the repeated CV estimator is recommended for general use. Currently, cross-validation is widely accepted in data analysis and machine learning, and serves as a standard procedure for performance estimation and model selection.

There are some new CVs for time series in recent years. Bergmeir [4] proposed blocked cross-validation (BCV) in evaluating prediction accuracy. Opsomer [17] found that cross-validation will fail when the correlation between errors of time series exists. To solve the correlation, three new CVs called modified cross-validation (MCV), partitioned cross-validation (PCV) and hv-blocked cross-validation (hvBCV) were presented [7,19].

How to measure the generalization error is crucial for comparing time series models because different measurements may provide opposite results, e.g., models with low mean absolute error could have large mean relative error. Salzberg [23] proposed using k-fold CV followed by appropriate hypothesis test to compare models rather than the average accuracy. Many subsequent studies about comparing algorithms are in the schema of cross-validation and hypothesis test (CV & HT) [8]. The variance of error estimation is needed in most hypothesis tests. In addition, Rodriguez [21,22] compared the estimator for different folds of CV and concluded that if the aim is to compare classifiers with similar bias, 2-fold CV is advocated because it has the lowest variance. Therefore, the variance of estimator is very important for comparing models.

The variance of errors is usually estimated before hypothesis tests. On the one hand, the classical variance estimator would be grossly underestimated due to the overlap between training and testing sets [2,3,23]. On the other hand, if series autocorrelation is present, the test error will also be underestimated, but CV is not able to detect this [19]. Existing CVs do not solve above problems at the same time.

This paper aims to design an effective error estimation method for time series models. Considering the periodicity, overlapping or correlation of series, M-CV with Markov property is proposed. Its randomness and independence could overcome the above problems, and the equiprobability and representativeness could balance CV subsets. Furthermore, its low variance could promote the error estimation. These characters ensure that M-CV could provide an effective and accurate estimation of generalization error.

The paper is organized as follows. In Section 2 three criteria are summarized for model evaluation of time series. Based on them, M-CV methodology is proposed. In Sections 3 and 4, some sound properties of M-CV are subsequently illustrated and it is compared with other CVs in theory and experiments. Section 5 concludes.

2. M-CV methodology

2.1. Time series model

This paper focuses on time series smoothing model. Time series smoothing or fitting is a basic representation technique which can be used for distance measures, time series compression, clustering and so on [24].

For a common time series $S = \{y_{t_i}\}, (i = 1, 2, \cdots, n)$, the conventional time series smoothing aims to estimate a function $f(\cdot)$ which could reflect the real series to some extent. It is essentially single-input regression. Time series could be expressed as: $y_{t_i} = f(t_i) + \epsilon_i$, where ϵ_i denotes noise component.

2.2. CV criteria

2.2.1. Randomness of partition

Seasonal and cyclical components usually exist in time series. If series is partitioned periodically in CV procedure, models are likely to learn biased information and may produce inaccurate error estimation. This can be illustrated by the following example.

Fig. 1 shows monthly series of carbon dioxide content in Mauna Loa within 16 years (1965.1∼ 1980.12) [9]. Two subseries (series in April and October) and smoothed curves are plotted in Fig. 1. It can be observed that the original series has an obvious seasonal component. The values in April and October are peaks and valleys of series, respectively. Obviously, the two smoothed curves are biased for the whole series. Moreover, if a model is trained on peak points and tested on valley points, the prediction error will be overestimated. Thus periodic partition should be avoided. This can be achieved by the partition with randomness.

2.2.2. Independence of test errors

The variance of test errors is usually estimated by sample variance. However, if we do not take into account the error correlations due to the overlap between training or test sets, naive variance estimator will seriously underestimate the variance [2,20]. There is no overlap between test sets because each example of the original data set is used once and only once as a test example [2]. For most k-fold CVs, there are additional dependencies between training sets. An exception is 2-fold CV whose test errors are independent since the training sets do not overlap [1,2,8].

2.2.3. Independence between training set and test set

If a series is autocorrelated, the model is easily overfitted and the test error will be underestimated [19]. Independence can be assured by leaving a certain distance between training and test samples. In other words, if a sample appears in test set, all other correlated samples have to be removed from the training set to avoid overfitting on the sample. Thus CV partition on time series has to leave a certain distance to keep the independence between training set and test set [4,5,12].