A New Nonparametric Algorithm for Preprocessing Stochastic Data with Uncertainty

Ekaterina Chzhan[1*]

[1]Siberian Federal University, School of Information and Space and Technology, 660074 Krasnoyarsk, Russian Federation

Abstract. The article deals with the problem of modeling stochastic processes under uncertainty. The peculiarity of the processes under consideration is that the researcher does not have information about the mathematical structure of the object; the object is represented as a black box. The article proposes to use a nonparametric modeling algorithm based on a nonparametric estimate of the regression function on observations. To improve the accuracy of modeling, it is proposed to use an algorithm for generating training samples. The algorithm differs from the previous modification by the definition of essential variables. The results of computational experiments have shown the effectiveness of the proposed algorithms.

1 Introduction

The problem of data analysis and modeling of stochastic processes is relevant for many subject areas, including in the case of processing geoinformation data [1]. In the case when the mathematical structure of the object under study is unknown, one of the solutions can be the use of nonparametric estimates [2]. As it is known, the accuracy of nonparametric models depends on many factors. One of the key factors is the quality of the initial data [3]. In the original data there may be outliers due to measurement errors and missing data due to different discreteness of control [4]. These features can negatively affect the quality of modeling, and in some cases lead to unsatisfactory results. Another feature is the uneven sparse distribution of the sample elements in the space of input and output variables [5]. In this case, resampling algorithms [6] or bootstrap methods [7] can be used. If the sample sizes are small with a large space dimension, it is proposed to use methods for generating training samples. However, the previously proposed methods for preprocessing sparse samples cannot be applied in conditions of noisy data (in the presence of outliers), as well as in the presence of a large number of variables, including insignificant ones. Modification of existing algorithms for obtaining training samples will expand the conditions for the applicability of these methods and improve the accuracy of modeling.

The rest of the article is organized as follows. In Section 2, we present the problem statement. Section 3 contains a proposed algorithm our, and in Section 4 we analyze the results of our simulation study data. Discussion and conclusions are presented in Section 5.

* Corresponding author: echzhan@sfu-kras.ru

2 A problem statement

Consider the proposed process modeling scheme presented in Figure 1 [8]. The following designations are accepted here: $x(t)$ is an output process variable, $\boldsymbol{u}(t)$ is an input vector action that can be controlled and measured. Random action is $\xi(t)$, control blocks of input and output variables are G^u and G^x. Random noises $g^x(t)$, $g^u(t)$ have zero mathematical expectation and limited variance. Measurements of input and output variables form a sample of observations $\left\{\boldsymbol{u}_i, x_i, i=\overline{1,s}\right\}$, which is given to a data analysis block. This module contains data preprocessing algorithms for training sample generating.

The modeling task is not only to obtain an estimate $\hat{x}(t)$ of the output process variable $x(t)$, but also to process the initial data.

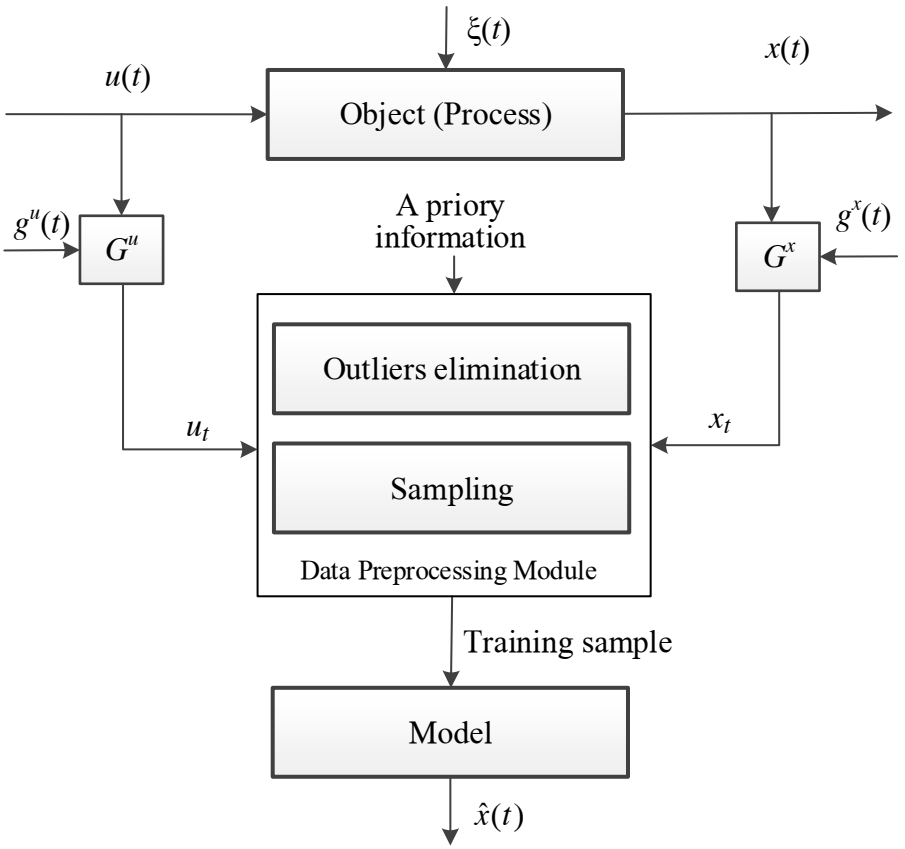

Fig. 1. A discrete-continuous process identification scheme.

3 A new modification of nonparametric algorithm for data preprocessing

As a model, we use the well-known nonparametric Nadaraya-Watson (NW) estimator [9, 10]:

$$\hat{x}(\boldsymbol{u}) = \frac{\sum_{i=1}^{s} x_i \prod_{v=1}^{m} K\left(\frac{u^v - u_i^v}{c_s^{u[v]}}\right)}{\sum_{i=1}^{s} \prod_{v=1}^{m} K\left(\frac{u^v - u_i^v}{c_s^{u[v]}}\right)}, \quad (1)$$

where $\boldsymbol{u} = (u_1, u_1, \ldots, u_m)$, c_s is the bandwidth or smoothing parameter, which is unique for every variable, $K\left(\frac{u^v - u_i^v}{c_s^{u[v]}}\right)$ is a kernel function.

The estimate belongs to the class of local approximation, the accuracy depends on the number of observations that are in the c_s-neighborhood of the point under consideration. An algorithm for generating additional observations was proposed earlier [11]. However, if there are outliers in the sample, the result will be unsatisfactory. In addition, all variables were previously used to calculate the estimate (1) in [12]. As it is known, the inclusion of insignificant variables in the model can lead to a decrease in accuracy. Therefore, it is proposed to identify significant variables and only use them both for calculating the estimate and in the case of generating a new training sample The paper proposes to modify the algorithm as follows:

1. Outliers identification and treatment using nonparametric method [13].
2. Determination of smoothing parameters $c_s^{u[1]}, c_s^{u[2]}, \ldots, c_s^{u[m]}$ by minimizing the criterion of compliance of the object's output $x(\boldsymbol{u})$ with the model $\hat{x}(\boldsymbol{u})$ in the cross-validation mode:

$$I\left(c_s^{u[1]}, c_s^{u[2]}, \ldots, c_s^{u[m]}\right) = \\ = \sum_{l=1}^{s} \left(\hat{x}\left(u_{1,l}, u_{2,l}, \ldots, u_{m,l}\right) - x_l\right)^2 = \min_{c_s^{u[1]}, c_s^{u[2]}, \ldots, c_s^{u[m]}}, l \neq i \quad (2)$$

   where the index i is the observation number in (1). The Nelder-Mead multidimensional optimization method can be used as an optimization algorithm. This method allows you to work with nonsmooth and noisy functions.
3. Among all the found coefficients $c_s^{u[1]}, c_s^{u[2]}, \ldots, c_s^{u[m]}$, select the maximum value. This coefficient corresponds to the least significant component $\boldsymbol{u}$.
4. Calculate $\hat{x}(\boldsymbol{u})$ due to (1) excluding the factor $\Phi(\cdot)$ for which the optimal value of the smoothing parameter is maximum.
5. Calculate the value of the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{s} \sum_{i=1}^{s} \left(\hat{x}_i - x_i\right)^2 \quad (3)$$

6. The steps 2 – 5 will be repeated until $\mathrm{MSE}_n \leq \mathrm{MSE}_{n-1}$. Thus, all significant variables will be found. Let there be $k \leq m$ variables.
7. Calculate the average number λ of sample elements that are found in the c_s neighborhood of each sample element.
8. In the vicinity of observations $\dot{u}$, for which the number of neighbors is smaller than the sample mean λ, additional artificial sample elements are generated due to the formula:

$$\tilde{u}_{li}^j = \dot{u}_i^j + \psi_l^j c_s^{u[l]}, \; j = \overline{1,k}, i = \overline{1,s'}, l = \overline{1,v_i}, \quad (4)$$

where $\psi_l^j \neq 0$ is a random variable distributed according to a uniform law in the interval [-1; 1], $\dot{u}$ is the value of the input variables of the element, in the vicinity of which the values $\tilde{u}_{li}^j$ are generated.

9. For the generated items, instead of the object output we calculate estimation (1) with only significant components of vector $\boldsymbol{u}$. The generated elements and the original sample constitute a new training sample $\left\{\tilde{\boldsymbol{u}}_i, \tilde{x}_i, i = \overline{1, s_1}\right\}, s_1 \geq s$.

We obtain the Nadaraya-Watson kernel estimator (NWN) with new training sample $\left\{\tilde{\boldsymbol{u}}_i, \tilde{x}_i, i = \overline{1, s_1}\right\}$ as follows:

$$\hat{x}_{NWN}(\boldsymbol{u}) = \frac{\sum_{i=1}^{s_1} \tilde{x}_i \prod_{v=1}^{k} \Phi\left(\frac{u^v - \tilde{u}_i^v}{c_s^{u[v]}}\right)}{\sum_{i=1}^{s_1} \prod_{v=1}^{k} \Phi\left(\frac{u^v - \tilde{u}_i^v}{c_s^{u[v]}}\right)} \tag{5}$$

4 Simulation results

A simulation study was conducted to compare the performances of the estimators with the classical Nadaraya-Watson estimators. For the simulation, we used the following function:

$$x(u) = 7u_1 + 6\sqrt{u}_2 - 8u_3 + 2u_4^2 + 0.1u_5 + \xi \tag{6}$$

where the $u_l, l = \overline{1,5}$ were drawn from a uniform distribution based on the interval [0, 1]. The ξ have a normal distribution with 0 mean and 0.1 variance. In this way, we generated samples of size 200, 500 and 1000. Samples were sparse with outliers.

As a kernel function we used Epanechnikov kernel function [12]:

$$K(z) = \begin{cases} 3\left(1 - z^2\right)/4, \text{if } |z| \leq 1, \\ 0, \text{else} \end{cases} \tag{6}$$

For each sample, we computed the values of mean percentage absolute error (MAPE) related to the kernel estimations: standard Nadaraya-Watson estimation (1) NW, estimation NWS (1) in case of the training sample obtained with method described in [11] and estimation NWN (5). The MAPE values of the kernel estimators which are obtained using the Epanechnikov kernel functions (6) are given in Table 1.

Table 1. MAPE values of the estimations

s	NW	NWS	NWN
200	39.3	19.5	18.8
500	28.9	18.9	18.1
1000	24.3	18.1	17.3

As seen from Table 1, for all sample sizes, the kernel estimators NWN in case of usage the training sample obtained with the proposed algorithm of data preprocessing have smaller MAPE values than the NWS kernel estimator. In each case, it is seen that NWS has the best performance.

5 Conclusion

In the paper, we have studied the new modification of data preprocessing algorithm. The results of the simulation study, which was performed to evaluate the performances of the kernel estimators considered, showed that the use of training samples obtained with the proposed algorithm has improved the accuracy of modeling.

Acknowledgments: the study was supported by a grant from the President of the Russian Federation for state support of young scientists MK-763.2020.9.

References

1. D. Zhang, Y. Zhang, H. Li. Ad. in Marine Sc. **1**, 013 (2009)
2. A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, (2008)
3. B. Busygin, S. Nikulin. *Specialized geoinformation system RAPID: features, structure, tasks* in 14th EAGE International Conference on Geoinformatics-Theoretical and Applied Aspects, **1** (2015)
4. S. Efromovich. *Missing and modified data in nonparametric estimation: with R examples*. CRC Press (2018)
5. J. Chen, D. Zhang, W. Zhou, Z. Chen, H. Li. (2020). Str. and Infrastr. Eng. (2020)
6. Ł. Smaga. Com. in St.-Sim. and Comp. **46**, 10 (2017)
7. M.R. Chernick, W. González-Manteiga, R.M. Crujeiras, E.B. Barrios. *Bootstrap methods* (2011)
8. V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, (2013)
9. W. Hardle. *Applied Nonparametric Regression*, Cambridge, New Rochelle (1990)
10. S. Demir, O. Toktamis. On the adaptive Nadaraya-Watson kernel regression estimators, H. J. of M. and St. **39**, 3 (2010)
11. A.V. Medvedev, E.A. Chzhan. *On nonparametric modelling of multidimensional noninertial systems with delay*, B. of the S. Ur. St. Un. Ser. M. M., Pr. & Com. S. **10**, 2 (2017)
12. M. A. Denisov, E. A. Chzhan, A. A. Korneeva. *Non-parametric approach for preliminary processing of earth remote sensing data* in E3S Web of Conferences, **75**, 01015 (2019)
13. G.S. Watson. *Smooth regression analysis*, In. J. of St., **26**, 4 (1964)