International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-2, December 2020

A Nonparametric Algorithm for Data Preprocessing and Modeling Multidimensional Objects with Delay

Ekaterina Chzhan

Abstract: ***The paper devotes to modeling tasks of multidimensional inertialess objects with delay (MIOD). The description of identification scheme of MIOD is determined. The proposed identification scheme includes not only blocks of a process and a model but also a data preprocessing block to improve modelling accuracy. A new method of data preprocessing which includes outliers detection and sparsity filling is proposed. It allows generating new training samples based on initial data that is obtained by measurement of input and output variables of the process. A software package is developed to conduct computer experiments. The results of the study show that the proposed algorithms are universal and can be applied to simulate various objects that are described with liner, nonlinear algebraic and nonlinear transcendental mathematical equations. Computational experiments have shown satisfactory accuracy of the algorithms. Proposed algorithms can be used in modeling and control tasks for inertialess objects in various areas of industry such as metallurgy, petrochemicals and etc.***

Keywords: ***non-parametric algorithm, multidimensional process, data analysis, data preprocessing.***

I. INTRODUCTION

The problem of identification of multidimensional discrete-continuous systems is relevant today [1]. In the majority of works [2-5] aspects of dynamic systems modeling are considered. However, in practice, a situation may arise when the process is dynamic, but due to the large value of discreteness of control, we are forced to consider it as inertialess processes with delay. For example, the process time constant can be 2 – 3 minutes, while the measurement resolution is 30 – 40 minutes. In this case, there is no way to track the change in variables.

In the works [6-7] aspects of constructing parametric models of such systems in conditions of a large amount of a priori information are considered. Researchers have information about the type of mathematical description of the object. In practice, we do not always have an opportunity to determine the type of this dependence and then it is advisable to use nonparametric methods [8-10].

Revised Manuscript Received on November 25, 2020.

Ekaterina Chzhan*, Assistant Professor, School of Space and Information Technology, Siberian Federal University, Krasnoyarsk, Russian Federation. Email: echzhan@sfu-kras.ru

On this way, nonparametric methods, neural networks in particular, have found wide application [11]. The peculiarity of these methods is that training requires a large amount of initial data, and the required size of training samples increases significantly with increasing dimension. In some cases, it is necessary to work in conditions of small samples. For example, in the work [12], the process of steel melting was considered, where the dimension of the vector of input variables was 16 with a sample size of 189 measurements.

Using small samples leads to the problem of preliminary data analysis. The accuracy of the methods used directly depends on the quality of the initial data [13-14]. Thus, samples can be sparse. A number of works are devoted to working with spatial data samples [15]. Moreover, samples may contains outliers and omissions [16].

Thus, the paper considers new nonparametric methods for data preprocessing and modeling. A combination of these methods will improve the accuracy of solving the identification problem as a whole. This research work is discussed as follows. Section II and III present a problem statement and nonparametrics methods, developed system is described in section IV. The results are discussed in section V and the conclusion is given in section VI.

II. IDENTIFICATION PROBLEM STATEMENT

Many technological objects dominated by discrete-continuous processes with delay can be represented in a simplified form in the form of a diagram shown in Figure 1 [17]. The identification scheme includes blocks of a process, data preprocessing and a model. We identify the concepts of an object and the process taking place in it. In Figure 1 the notation is the following: u(t) is controlled input action, μ(t) is measured but uncontrolled input action, ξ(t) is random action, Gu, Gμ, Gx are control blocks of input and output variables with random interference gu(t), gμ(t), gx(t), x(t) is an output process variable, is an estimation of the output process variable (model output) [17]. Also, the object can be affected by variables λ(t) that have a significant impact, but cannot be controlled. Measurements of input and output variables are delivered to the block "Data Preprocessing", after processing we have a sample of observations.

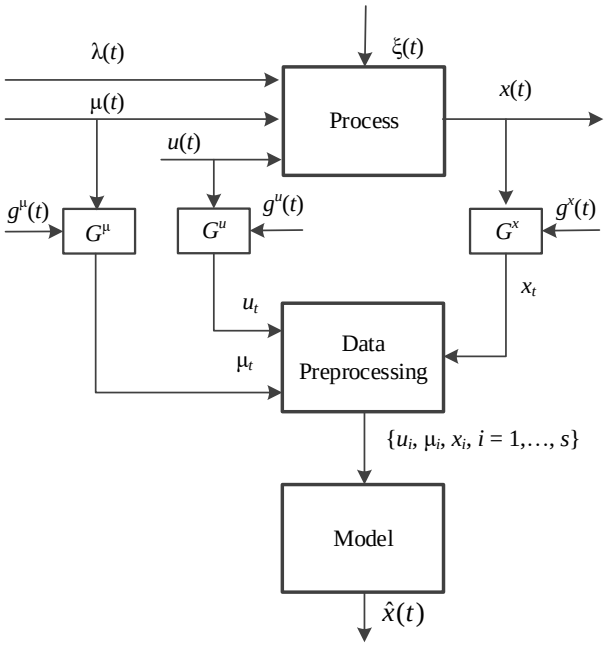

Fig. 1.An identification scheme of discrete continuous objects.

The identification task is to construct a model of the process under study. Next, the description of the developed algorithms of modeling and data preprocessing is described in more detail.

III. NONPARAMETRIC MODELING ALGORITHMS

Let's us consider a modification of nonparametric Nadaray-Watson estimate:

$$\hat{x}(u,\mu)=\frac{\sum_{i=1}^{s}x_i\prod_{j=1}^{m}\Phi\left(c_s^{-1}\left(u^j-u_i^j\right)\right)\prod_{j=1}^{k}\Phi\left(c_s^{-1}\left(\mu^j-\mu_i^j\right)\right)}{\sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(c_s^{-1}\left(u^j-u_i^j\right)\right)\prod_{j=1}^{k}\Phi\left(c_s^{-1}\left(\mu^j-\mu_i^j\right)\right)}, \quad (1)$$

where $\Phi(z), z=c_s^{-1}\left(u-u_i\right)$ is a kernel function, c_s is a bandwidth parameter, k is a number of input uncontrolled variables $\mu(t)$, m is a number of input controlled variables $u(t)$.

The estimation (1) includes not only controlled variables $u(t)$, but also uncontrolled variables $\mu(t)$. The value of the bandwidth parameter is found by minifying the quadric criterion of the correspondence between the object and the model with cross validation:

$$R(c_s)=\sum_{l=1}^{s}\left(x(u_l,\mu_l)-\right.$$
$$\left.-\frac{\sum_{i=1}^{s}x_i\prod_{j=1}^{m}\Phi\left(c_s^{-1}\left(u_l^j-u_i^j\right)\right)\prod_{j=1}^{k}\Phi\left(c_s^{-1}\left(\mu_l^j-\mu_i^j\right)\right)}{\sum_{i=1}^{s}\prod_{j=1}^{m}\Phi\left(c_s^{-1}\left(u_l^j-u_i^j\right)\right)\prod_{j=1}^{k}\Phi\left(c_s^{-1}\left(\mu_l^j-\mu_i^j\right)\right)}\right)^2=$$
$$=\min_{c_s},\ l\neq i. \quad (2)$$

The following estimates can be used as a kernel function:

$$\Phi(z)=\begin{cases}1-|z|, \text{if } |z|\le 1,\\ 0, \text{if } |z|>1;\end{cases} \quad (3)$$

$$\Phi(z)=\begin{cases}0{,}335-0{,}067(z)^2, \text{if } (z)^2\le 5,\\ 0, \text{if } (z)^2>5.\end{cases} \quad (4)$$

where $z=c_s^{-1}\left(u^j-u_i^j\right)$.

It should be noted that in order to improve the accuracy of modeling, it is proposed to use an algorithm for filling in the gaps. The main idea is that the missing values of the output variables are replaced by a nonparametric estimate of the form (1). Also, if the initial data has sparsity structure, it is advisable to use the algorithm [17] for generating additional values in those regions of the space of input and output variables where these sparsity subregions exist. But, if initial data sets have outliers, the use of such an algorithm may decrease in modeling accuracy. So, it is necessary to use trimming methods [18]. In proposed trimming method the main idea is to remove the points for which the residual between the model (1) and the object output is the most significant. Thus, the sequence of the nonparametric algorithm consists in filling in the missing values, removing outliers and generating additional sample elements in sparsity subregions.

IV. PROPOSED SYSTEM FOR MULTIDIMENSIONAL OBJECTS MODELING

A proposed system for multidimensional objects modeling implemented in the Visual Studio environment in the C # language was developed to conduct experiments on modeling multidimensional processes with delay. A scheme of the system is shown in Figure 2.

Retrieval Number: 100.1/ijeat.A19301010120
DOI:10.35940/ijeat.A1930.1210220
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

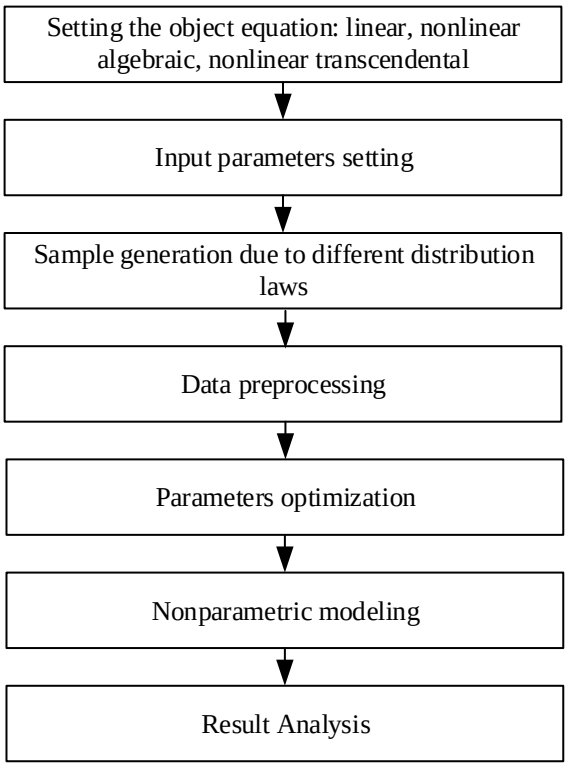

Fig. 2. Module blocks of the proposed system

At the initial stage, the user sets the dimension of the object and its mathematical description. Then it is necessary to define the input parameters of the simulation and generate a sample of observations. The values of the input variables are generated in accordance with the selected distribution law using the P-generator of random numbers [9]. The bandwidth parameter cs is adjusted using optimization algorithms. The key block is the construction of nonparametric models [10]. To conduct research, new examining samples are generated for the object under consideration.

V. RESULT AND DISCUSSION

The performance of proposed system based on nonparametric algorithms for MIOD is assessed through computer simulations. In series of computer experiments, let objects have the following mathematical descriptions:

$$x_1 = 1u_1 + 2u_2 - 0.7u_3 + u_4 - 2u_5 + 0.3u_6 - 1.6u_7, \quad (5)$$

$$x_2 = (u_1 + 0.3u_2)^{1.5} + 4u_3 + 0.6u_2u_3^2 - \\ -0.5u_4u_5^{0.5} + 2(u_6 - u_7)^3, \quad (6)$$

$$x_3 = 3^{u_1} + 0.5u_2^2 - 2\sin u_3 + u_4^{0.5} - \\ -0.3\log(u_5) + 0.4u_4^2u_5 + 0.3e^{u_6} - u_7. \quad (7)$$

According to research [19] we investigate objects of three main types: liner, nonlinear algebraic and nonlinear transcendental. The initial samples are generated in such a way that they contain sparsity subregions and outliers.

In computational experiments, the accuracy of the algorithm is estimated using the simulation error:

$$W = \sqrt{s^{-1}\sum_{i=1}^{s}(x_{si} - x_i)^2 \Big/ (s-1)^{-1}\sum_{i=1}^{s}(x_i - \hat{m}_x)^2}, \quad (8)$$

where $\hat{m}_x$ is mathematical expectation of x.

We use an initial data sample and a sample preprocessed using the proposed nonparametric algorithm as a training sample. We calculate errors (8) in both cases: W_1 and W_2 respectively. The error value is averaged over a series of 100 experiments.

Table-I: Modeling results

Object	Training sample		Examining sample	
	W_1	W_2	W_1	W_2
x_1 (5)	0.135	0.083	0.123	0.092
x_2 (6)	0.142	0.097	0.154	0.109
x_3 (7)	0.221	0.101	0.203	0.182

The research results show that the use of data processing algorithms make it possible to improve the accuracy of modeling up to 2 times.

VI. CONCLUSION

In this paper, nonparametric algorithms for modelling and data preprocessing is designed and investigated. The proposed modeling system includes algorithm for new training samples generating. The experimental result demonstrates that the proposed system provides excellent performance with high modeling accuracy for different objects.

ACKNOWLEDGMENT

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. L. Lizhen, L. Xu, Zh. Lin, "Stability and stabilisation of linear multidimensional discrete systems in the frequency domain", International Journal of Control, vol. 86, no 11, 2013, pp. 1969-1989.
2. P. Van Geert, "Dynamic systems approaches and modeling of developmental processes", Handbook of developmental psychology 640672, 2003.
3. N. Simidjievski, L. Todorovski, S. Džeroski, "Modeling dynamic systems with efficient ensembles of process-based models", PloS one, vol. 11, no 4, 2016, e0153507.
4. L. Xu, X.R. Li, Y. Liang, Z. Duan, "Constrained dynamic systems: Generalized modeling and state estimation", IEEE Transactions on Aerospace and Electronic Systems, vol. 53, no 5, 2017, pp. 2594-2609.
5. C. A. Kluever, "Dynamic systems: modeling, simulation, and control", John Wiley & Sons, 2020.
6. E. Walter, L. Pronzato, "Identification of parametric models", Communications and control engineering, vol. 8, 1997.
7. M. Myllymäki, M. Kuronen, T. Mrkvička, "Testing global and local dependence of point patterns on covariates in parametric models", Spatial Statistics, 2020 100436.
8. P. Orbanz, Y. W. Teh, "Bayesian Nonparametric Models", Encyclopedia of machine learning, vol. 1, 2010.
9. A. Chiuso, G. Pillonetto, "Bayesian and nonparametric methods for system identification and model selection", European Control Conference (ECC) IEEE, 2014, pp. 2376-2381.
10. J. L. Horowitz, "Applied nonparametric instrumental variables estimation", Econometrica, vol. 79, no 2, 2011, pp. 347-394.
11. J. A. Suykens, J. P. Vandewalle, B.L. de Moor, "Artificial neural networks for modelling and control of non-linear systems", Springer Science & Business Media, 2012.

12. A.V. Medvedev, M.E. Kornet, E.A. Chzhan, “Nonparametric modeling of oxygen-converter processes”, Steel in Translation, vol. 46, no 12, 2016, pp. 855-859
13. S. García, J. Luengo, F. Herrera, “Data preprocessing in data mining”, Cham, Switzerland : Springer International Publishing, 2015, pp. 195-243.
14. M. Huebner, S. le Cessie, C.O. Schmidt, W. Vach, “A contemporary conceptual framework for initial data analysis”, Obs Stud, vol. 4, 2018, pp. 171-92.
15. J.M. Loh, M.L. Stein, “Spatial Bootstrap with Increasing Observations in a Fixed Domain,” Statistica Sinica, vol. 18, 2008, pp. 667-688.
16. X. Su, C.L. Tsai, “Outlier detection”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no 3, 2011, pp. 261-268.
17. A.V. Medvedev, E.A. Chzhan, “On nonparametric modelling of multidimensional noninertial systems with delay”, Bulletin of the South Ural State University. Ser. Mathematical Modelling, Programming & Computer Software, vol. 10, no 2, 2017, pp. 124-136.
18. S. K. Kwak, J. H. Kim, “Statistical data preparation: management of missing values and outliers”, Korean journal of anesthesiology, vol. 70, no 4, 2017, pp. 407-411.
19. G. L. Shevlyakov, N.O. Vilchevski, “Robustness in Data Analysis: criteria and methods”, Walter de Gruyter, 2011, 309 p.

AUTHOR PROFILE

Dr. Chzhan Ekaterina is currently working as an Associate Professor in School of Space and Information Technology, Siberian Federal University, Krasnoyarsk, Russian Federation. Dr. Chzhan has received her Ph.D. in System Analysis. Her areas of interests are nonparametric modeling, data analysis and control systems.