FINDING SIGNIFICANT VARIABLES IN THE PROBLEMS OF MODELLING LESS MEMORY PROCESSES

E.D. Mihov
Siberian Federal University
Krasnoyarsk, RUSSIA
e-mail: edmihov@mail.ru

Abstract

The report is devoted to the task of identifying significant variables. There are many algorithms for the selection of significant variables, but they all have their weaknesses. A new algorithm for extracting significant variables based on non-parametric algorithms has been proposed. The result of the proposed algorithm is shown in the report.
Keywords: data science, memoryless process, significant variable

1 Introduction

The identification of many stochastic objects is often reduced to the identification of static systems [3]. The general scheme of the discrete-continuous process under study is presented in the Figure 1:

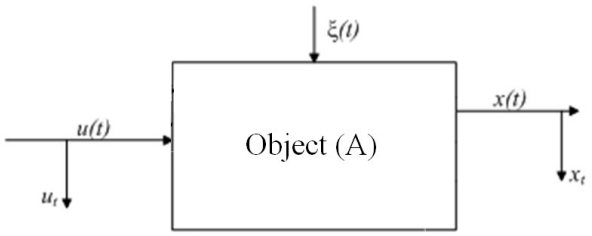

Figure 1: The general scheme of the process under study

In figure 1, the following notations: A is the unknown operator of the object, $x(t)$ is the output characteristic of the process, $\vec{u}$ is the vector of control actions, $\xi(t)$ is the interference (interference affects the accuracy of the reading of the output variables, for example, if the interference is 5%, then this means that the read output variable contains an error of up to 5%.).

The need to find significant input variables that affect $x(t)$ often arises in the study of such processes. The process of diagnosing diseases can be cited as an example.

If a technique for diagnosing any disease is not developed, then it is important to understand which factors will be of greater importance for diagnosis.

This is the reason to recognize the relevance of research in this area.

There are two ways to assess the significance of variables, direct and indirect [2].

A direct way to assess significantly variables is discussed in the report. It is important to note that the article discusses inertialess processes. For other types of processes, the allocation of significant variables may require additional actions.

The direct one is that it is necessary to find the vector of variables $\vec{u} = (u_1, u_2, \ldots, u_n)$, $R(\vec{u}, \vec{c}_s) \to 0, R(\vec{u}, \vec{c}_s)$ – average model error.

The task is to select m ($m < n$) of the n variables. Several algorithms exist to accomplish this task.

2 Del algorithm

The researcher should exclude u_1 and calculate $R(u_2, \ldots, u_n)$ the model constructed using variables $u_2, \ldots, u_n$ where $R(u_2, \ldots, u_n)$ is the average classification error. Then $R(u_2, \ldots, u_n)$ excluded in the same way and $R(u_1, u_3, \ldots, u_n), \ldots, R(u_1, \ldots, u_{n-1})$ are calculated. The variable with the least degree of significantly is by the rule: $maxR(u_1, \ldots, u_i, \ldots, u_n) \to minI(u_i), u_i$ is the variable with the least degree of significantly. This variable is excluded from the modelling process. $n-1$ variables remains after that. The algorithm must be repeated until there are m significant variables left.

The number of iterations (L) performed for the selection of significant variables is calculated by the formula (1) with this approach.

$$L = n + (n-1) + (n-2) + \ldots + (m+1) = \sum_{i=1}^{n-m} (n-i) \tag{1}$$

3 Algorithm Ad

According to this algorithm is necessary estimate $R(u_i)$ for every u_i . The significant variable is selected by the rule: $minR(\cdot) \to maxI(\cdot)$, where R is the error of variable $(\cdot)$ based model, I is the significant of variable. Thus the first significant variable will be found. Variables $u_1, \ldots, u_{i-1}, u_{i+1}, \ldots u_n$ are alternately added to the variable found. As a result, the researcher receives the appropriate sets of variables $(u_i, u_1), \ldots, (u_i, u_n)$ on the basis of which he calculates $R(u_i, u_1), \ldots, R(u_i, u_n)$. The essential set of variables is selected by the above rule again. This operation must be repeated until a set of n variables is typed. The number of iterations done to select significant variables is equal to the number of iterations in the Del algorithm for this approach.

4 AdDel Algorithm

The considered algorithms of selection of essential variables are included in the class of the so-called "greedy" algorithms. The problem with such algorithms is that when

an optimal solution is obtained at each step, they do not provide a global optimum.

To partially eliminate the shortcomings of these algorithms, use the algorithm Ad-Del. According to this algorithm, the researcher first finds a_1 the significant variables using the Ad algorithm. Next, the researcher using the algorithm Del excludes $a_2(a_2 < a_1)$ variables from the selected ones. This algorithm is repeated until a set consisting of n variables is found.

It was proved that the presented algorithms for finding significant variables do not always lead to a satisfactory result [1].

5 Algorithm for finding significant variables based on the setting of the blur coefficient parameter

A new method for finding significant variables was proposed - an algorithm for evaluating significantly of variables based on the setting of the blur coefficient parameter.

Before using the algorithm for evaluating significantly of variables based on the setting of the blur coefficient parameter, it is necessary to center and normalize the elements of the vector u from the training set.

The nonparametric evaluation of the regression function from observations is (2) [3]:

$$x_s(\vec{u}) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{n} \Phi(\frac{u_j - u_{ji}}{c_{sj}})}{\sum_{i=1}^{s} \prod_{j=1}^{n} \Phi(\frac{u_j - u_{ji}}{c_{sj}})}, \tag{2}$$

$x_s(\vec{u})$ is a non-parametric function estimate $x_s(\vec{u})$ in the point $\vec{u}$; $\Phi(\cdot)$ is bell-shaped function; s is the sample size of observations; $\vec{c}_s$ is blur parameter vector.

The corresponding component of the vector $\vec{c}_s$ is associated with each component of the vector $\vec{u}$ as can be seen from (2). Further, based on the available training set, it is necessary to find the optimal $\vec{c}^*_{s1}, \vec{c}^*_{s1} \ldots \vec{c}^*_{sn}$ from the minimum condition (3):

$$\sigma(\vec{c}_s) = \sqrt{\frac{1}{s} \sum_{i=1}^{s} (x_s(\vec{u}_i, \vec{c}_s) - x_k)^2} \to min, \tag{3}$$

and $k \neq i$.

After finding the vector c^*_s, sorting vector elements c^*_s from lowest to highest to highest should be produced. The chain of inequalities will turn out this way, for example: $|c_{s2} * | < |c_{s9} * | < |c_{s1} * | < \ldots < |c_{s4} * | < |c_{s3} * |$. The component of the vector $\vec{u}$ for which the corresponding element of the vector $\vec{c}_s*$ was the largest, is the candidate for the exception from the nonparametric estimation, as the least significant.

Computational experiment. The object under study has 10 input variables and 1 output variable. The impact of interference is 5%. The sample size of observations is 1000. Optimization is performed using the Neddler – Midd algorithm (a deformable polyhedron). The output variable is affected by two non-essential variables.

The object is described by the formula (4):

$$x(\vec{u}) = 3.62u_1 + 3.69u_2 + 3.79u_3 + 0.75u_4 + 3.73u_5) + \quad (4)$$
$$3.61u_6 + 3.79u_7 + 3.78u_8 + 0.62u_9 + 3.61u_{10}$$

Note that the formula describing the object is unknown to the researcher. This formula is used only for generating the sample.

The variables u_4 and u_9 are non-informative. Therefore, they should be excluded according to the rule formulated above.

The results of the calculations are presented below.

At the tact 1 was found so components of vector:

$\vec{c}_s = \{0.59; 0.55; 0.53; 1.26; 0.87; 0.56; 0.70; 0.69; 2.72; 0.51\}$, modeling error = 5.5%.

At the tact 2 was found so components of vector:

$\vec{c}_s = \{0.83; 0.91; 0.43; 1.23; 0.90; 0.55; 0.53; 0.87; -; 0.75\}$, modeling error = 5.0%.

At the tact 3 was found so components of vector:

$\vec{c}_s = \{0.70; 0.64; 0.58; -; 0.63; 0.69; 0.67; 0.65; -; 0.57\}$, modeling error = 4.3%.

At the tact 4 was found so components of vector:

$\vec{c}_s = \{-; 0.76; 0.48; -; 0.69; 0.67; 0.59; 0.72; -; 0.58\}$, modeling error = 4.3%.

The symbol "-" denotes the absence of the components of the vector u when describing the process under study. Non-essential variables were excluded in the first two tacts as shown in experiment. The elimination of the significant variable u1, in the third tact, led to an increase in the simulation error. Thus, the above method of finding significant variables in the problems of identification seems encouraging. This is confirmed by the results of numerous computational experiments.

6 Conclusion

A method for evaluating the significantly of process variables has been proposed. The results of using this method, which prove its effectiveness, were presented. The complexity of the method lies in optimizing the vector of blur factors.

This variable selection method allows reducing the number of modelling errors and eliminating non-essential variables. Studies can be applied in medicine for the diagnosis of diseases or in little-studied technical processes.

This work was supported by the Ministry of Education and Science of the Russian Federation in the framework of the Federal target program «Research and development on priority directions of development of the scientific-technological complex of Russia for 2014-2020» (agreement № 14.578.21.0247, unique ID project RFMEFI57817X0247)

References

[1] Francuz A.G. (1964). Adaptation and training in automatic systems. *Izv. AN SSSR. Ser. Tekhn. Kibernetika.*. Vol. **4**, pp. 68-77.

[2] Zagorujko N.G. (2013). *Cognitive data analysis*. Geo, Novosibirsk.

[3] Tsypkin Ya.Z.(1968). *Adaptation and training in automatic systems*. Nauka, Moskva.