Bootstrap and Counter-Bootstrap approaches for formation of the cortege of Informative indicators by Results of Measurements

M V Artemenko¹, I E Chernetskaia², N M Kalugina³ and E N Shchekina¹

¹Department of biomedical engineering, South-West State University, 94, 50 years old of October, Kursk, 305040, Russia
²Department of computer facilities, South-West State University, 94, 50 years old of October, Kursk, 305040, Russia
³Department of Space instrument making and communication systems, South-West State University, 94, 50 years old of October, Kursk, 305040, Russia

E-mail: artem1962@mail.ru

Abstract. This article describes the solution of the actual problem of the productive formation of a cortege of informative measured features of the object of observation and / or control using author’s algorithms for the use of bootstraps and counter-bootstraps technologies for processing the results of measurements of various states of the object on the basis of different volumes of the training sample. The work that is presented in this paper considers aggregation by specific indicators of informative capacity by linear, majority, logical and "greedy" methods, applied both individually and integrally. The results of the computational experiment are discussed, and in conclusion is drawn that the application of the proposed methods contributes to an increase in the efficiency of classification of the states of the object from the results of measurements.

Keywords: object of research, the information capacity of measured characteristics, bootstrap method, self-organization modeling.

1. Introduction

The development of computing and intellectual resources of modern decision support tools in scientific research and industry allows applying the achievements in this field for the analysis of complex and open systems by conducting adequate simulation. The simulation model implies a certain architecture and mathematical apparatus that describes the interaction of model elements, the functioning of internal and external interfaces. In the general case, the cybernetic simulation model describes the behavior of an object in the form of a certain multipolar network characterizing a set of values of the variously measured characteristics to ensure observability and controllability of the object.

At the initial stage of modeling, there are problems associated with complex technologies with inadequate reliability. The article [1] describes that the behavior of complex objects is characterized by a set of indicators ordered by a certain principle - a cortege, consequently, the result depends not
only on the measurement method, but also on the algorithm that determines the sequence of measurement of the indicators. Therefore, the problem of the formation of a cortege of informative features is actual.

2. Task
Research of various information sources shows that nowadays methods of forming a set of informative features of a modeling object are based on the application of the following methods and algorithms [2]: Full – a complete search of various combinations of characteristics to achieve an acceptable classification effect, Add and Del – consecutive addition and exclusion of features; AddDel – combining of algorithms Add and Del; Prob – each feature of the object is determined by weights and the procedures of the indicated algorithms are applied; fractal analysis - used for tensor data; Grad - is similar to the algorithm AddDel, but the inclusion and exclusion of characteristics in the final informative set is not "one by one", but "jointly".

Since the cortege implies a certain order, it is suggested that the criterion of informative sign of the characteristic be taken as the sorting criterion. The complexity of its definition is due to the fact that the mentioned algorithms function adequately in conditions of statistically representative volumes of training and examination samples. If the above condition is violated, the efficiency of the algorithms decreases drastically.

In this connection, there arises the problem of optimal formation of a tuple in conditions of different volumes of training material based on the application of measures characterizing the informative capacity of the feature from the point of view of certain classification possibilities for determining the state of the object in the process of performing the measurements. The application of the cortege will allow rationalizing the technology of the measuring process.

3. Theory
Let's allocate three main situations, characteristic for the decision of problems of classification, approximation and extrapolation:

1. The volume of data satisfies the statistical criteria of representativeness. In this case, standard methods for calculating informative capacity are used (Kullback’s method, Student’s t-test, classification errors, etc.), adapted to the subject domain in question.

2. The volume of information far exceeds the statistical threshold, which significantly increases the likelihood of obtaining false conclusions on typical statistical criteria. This situation arises quite often when measuring the characteristics of an object recorded with a high sampling frequency. In this case, it is recommended to reduce the sample size by applying convolution algorithms or reducing the dimensionality of the feature space. The drawback of this approach is the possible loss of information. In this regard, it is recommended that the sample be selected as a minimum for the information assessment method used in the future and gradually increase its volume to an acceptable value for the adequacy of the estimate. Let's call this approach a "counter-bootstrap".

3. The sample size is much smaller than the statistical threshold. In this case, it is recommended to use the bootstrap methodology for bringing the volume to the required level (considered, for example, in [4]) and apply the "law of small numbers" [5].

The theory of self-organization and the system approach allows us to combine the last two approaches to assess the informative capacity of the measured characteristics of an object, using the integration of properties of methods that require a large amount of information and allow obtaining acceptable results on the basis of a small number.

First approach is used in Rasch measurement [6], and second approach - in Group Method of Data Handling (GMDH) [7]. We denote as $Inf_{Rasch}(x_i)$ the quantitative evaluation of the informative capacity of the analyzed measurable or latent attribute.

In contrast to G. Rasch’s models, GMDH self-organization modeling algorithms, proposed by Ivakhnenko AG, work more adequately in conditions of small-volume samples, - especially at the stage of exploration analysis. Thus, we denote as $Inf_{MDH}(x_i)$ the quantitative evaluation of the informative capacity for GMDH.

As any artificial modification of the data sample size generally reduces the measure of confidence in them, it is recommended to use correction coefficient RRC (the reduction ratio of confidence):
where \( \text{por} \) is the threshold value \( RRC \), determined by the researcher (recommended values are \( 0 \leq \text{por} \leq 1 \)); \( d \) - the distance between the characteristic points of the original and modified sample in the plane of mathematical expectation and variance of the reduced values \( (0 \leq d \leq \sqrt{2}) \) (the reduced values of the elements of the sets of recorded characteristics are determined by normalizing the values of sets by linear transformations in the range \( [e, 1-e] \), \( e = \frac{1}{N} \) where \( N \) - initial sample size).

We denote these coefficients, respectively, as: \( RRC_b \) and \( RRC_{cb} \).

Guided by the hypothetical-deductive method of scientific research and the concepts of methodological invariant, we propose the following technology of calculations \( \text{Inf}_M \text{GDH}(x_j) \) [8]. At the first stage, the structural-parametric identification of the approximating polynomial \( \text{App}(X) \) [6] for each characteristic of the initial set \( \{X\} \) according to the equation:

\[
Y(Z) = A_0 + \sum_{k=1}^{K} \left( A_k \cdot \prod_{i=1}^{N} z_i^{p_{i,k}} \right),
\]

where: \( Z = \{z_1, z_2, \ldots, z_N\} \) is the set of arguments, \( Y(Z) \) is the response function, \( K \) is the number of terms in a polynomial, \( A_k, p_{i,k} \) - model’s parameters, \( N \) - number of arguments.

The identification procedure is repeated \( N \) times for each object state class \( w_l, l = 1, L \) \( (L \) is the number of classes) with the consequent formation of a sets \( \{Z\} = \{X\} - x_j \) and responses \( Y(Z) = x_j \).

As a result, sets of mathematical models for alternative classes \( w_l \):

\[
\{\text{App}(X)\}_{w_l}^{M_l}, \quad M_l \leq N, \quad M_l \neq 0, \quad R^2 [\text{App}(X)] \geq R^2_{lim},
\]

where \( R^2_{lim} \) is the threshold value of determination coefficient \( R^2 [\text{App}(X)] \). If, as a result of selection, an empty set is formed \( (M_l \neq 0) \), then models are successively included in it as the values of the determination coefficients decrease (the threshold value increases \( R^2_{lim} \)). The minimum volume of many mathematical models is recommended not less than 3.

After that, matrices are formed for each alternative class \( \{Rn\}_{w_l}^{M_l} \), the number of rows in which are equal, respectively, \( M_l \) columns - the number of indicators of the set \( \{X\} \), the value of matrix elements is calculated by successive application of equations:

At first, the fraction of influence of each term \( k \) in classes is determined by the equation:

\[
V_{k,w_l} = \frac{A_{k,l} \prod_{i=1}^{N} x_i^{p_{i,k,w_l}}}{\sum_{j=1}^{L} \left( A_j \prod_{i=1}^{N} x_i^{p_{i,j,w_l}} \right)},
\]

where \( \text{Median of ZZ values} \) determines the median of ZZ values as the most resistant to artifacts.

At second, for each argument included in the \( k \)-th term, the weight of the multiplier is calculated by the equation:

\[
M_{i,k,w_l} = \frac{|p_{i,k,w_l}| \ln (x_i)}{\sum_{j=1}^{N} |p_{i,j,w_l}| \ln (x_j)}.
\]

Next, the values of the additive-multiplicative influence of the exponent \( x_i \) are calculated on the response function \( Y(Z) \) for each alternative class:

\[
AM_{x_i,w_l} = 1 - \prod_{k=1}^{K} \left( 1 - V_{k,w_l} \cdot M_{i,k,w_l} \right).
\]

«Relative error of differences» (RED) \( \varepsilon \) is introduced (recommended values are \( 0.01 \leq \varepsilon < 0.1 \)) and the values of the magnitude of the multiplier effect are recalculated at \( [AM_{x_i,w_l}, \varepsilon] \) by the equation:
For each class, the characteristics of \(x_i\) are ordered as the values decrease \(AM_{x_i}^{w_i}, \varepsilon\) and 1 sets of characteristics for classes \(\{X\}^{w_i}\) and the corresponding sets of ranks \(\{Rn\}^{w_i}\) are formed. \(j = 1, \bar{N} -\) number of elements of sets of ranks, and elements are calculated by the equation:

\[
[AM_{x_i}^{w_i}, \varepsilon] = \begin{cases} 
AM_{x_i}^{w_i}, & \text{если } (1 - \varepsilon) \cdot AM_{x_i}^{w_i} < AM_{x_i}^{w_i} \leq (1 + \varepsilon) \cdot AM_{x_i}^{w_i}, \\
AM_{x_i}^{w_i}, & \text{otherwise}
\end{cases}
\]

(6)

For each class, the characteristics of \(x_i\) are ordered as the values decrease \(AM_{x_i}^{w_i}, \varepsilon\) and 1 sets of characteristics for classes \(\{X\}^{w_i}\) and the corresponding sets of ranks \(\{Rn\}^{w_i}\) are formed. \(j = 1, \bar{N} -\) number of elements of sets of ranks, and elements are calculated by the equation:

\[
Rn_j^{w_i} = \begin{cases} 
N, & \text{для } j = 1 \\
Rn_j^{w_i} - 1, & \text{для } (i \neq 1) & \text{и } ([AM_{x_i}^{w_i}, \varepsilon]_j \neq [AM_{x_i}^{w_i}, \varepsilon]_{j-1}) \\
Rn_j^{w_i}, & \text{для } (i \neq 1) & \text{и } ([AM_{x_i}^{w_i}, \varepsilon]_j = [AM_{x_i}^{w_i}, \varepsilon]_{j-1})
\end{cases}
\]

(7)

The set \(\{Rn\}^{w_i}\) forms the final set of informative characteristics for a certain volume \(N_f \leq N\), which includes elements \(X_{I_i} = (x_i / Inf MGDH(x_i)) / i = 1, \bar{N}\), imported from the source set \(\{X\}\) according to \(\{Rn\}^{w_i}\) sequentially, as the rank values decrease. The final informative capacity \(Inf MGDH(x_i)\) is calculated as:

\[
Inf MGDH(x_i) = \frac{\max_{j=1}^{\bar{N}}(Rn_j^{w_i})}{\max_{j=1}^{\bar{N}}(\max_{j=1}^{\bar{N}}(Rn_j^{w_i}))}.
\]

(8)

Using values of \(Inf Rasch(x_i)\) and \(Inf MGDH(x_i)\) for integrating of results of bootstrap and counter-bootstrap, we obtain the value of the integral information \(Inf(x_i)\) of the characteristics \(x_i\):

\[
Inf(x_i) = F(Inf Rasch(x_i), Inf MGDH(x_i), RRC_B, RRC_{cb}).
\]

Based on the research objectives, the data structure and the semantic load of the analyzed features of the modeling object, it is proposed to apply the following calculation equations \(Inf(x_i)\), where: \(a = RRC_B \cdot Inf Rasch(x_i), b = RRC_{cb} \cdot Inf MGDH(x_i)\):

- additive-weight aggregation is recommended in the case of noncollinear features, which were selected by models of G. Rasch and GMDH:

\[
Inf(x_i) = \frac{a + b}{RRC_B + RRC_{cb}}.
\]

(9)

- majoritarian aggregation is recommended when the characteristics are collinear, namely:

a) "careful" aggregation:

\[
Inf(x_i) = \min (a, b);
\]

(10)

b) "tense" aggregation:

\[
Inf(x_i) = \max (a, b);
\]

(11)

c) "greedy" aggregation, with the help of an operation «softmax» (according to R. Sutton [9]):

\[
Inf(x_i) = \max \left( \frac{\exp \left( \frac{a}{RRC_B + RRC_{cb}} \right)}{\exp \left( \frac{b}{RRC_B + RRC_{cb}} \right)}, \frac{\exp \left( \frac{b}{RRC_B + RRC_{cb}} \right)}{\exp \left( \frac{a}{RRC_B + RRC_{cb}} \right)} \right);
\]

(12)

d) Boolean aggregation:

\[
Inf(x_i) = \frac{1}{1 + \alpha} \cdot (a + b + \sqrt{a^2 + b^2} - 2 \cdot \alpha \cdot a \cdot b);
\]

(13)

where: \(\alpha\) is the parameter that takes a value from the interval (-1, 1).

Since the formation of a set of informative features according to the proposed methods is carried out according to the ordered values of information indicators, the problem of the formation of the sought-for cortege is solved.
4. Results
In order to study the possibilities of the control methods considered, a computer experiment was carried out: the object was in two alternative classes, \( n = 10; 30 \) and \( 100 \) signs were synthesized and signs were correlated and uncorrelated among themselves (determined by the value of the pair correlation \( R \)). Number of observations \( m = 500 \). To simulate the characteristics of the object was used randomization according to the uniform distribution law, the artificial error was carried out at the level of 20\%, and all attributes had a classification efficiency at the level of 0,65-0,7. Three options for the formation of training and examination samples were considered: \( m_1 < m \ (m_1 = 0.2m) \); \( m_1 > m \ (m_1 = 3m) \); \( m_1 \approx m \ (m_1 = 0.8m) \). The results of the computational experiment are given in Table 1.

| sample size | Collinearity R | aggregation type | number of informative signs \( n_1 \) | Classification efficiency |
|-------------|----------------|------------------|-------------------------------------|--------------------------|
| \( m_1 = 0.2m \) | 0.8 - 0.9 | additive-weight | 6; 21; 34 | 0.66; 0.66; 0.68 |
| | 0.8 - 0.9 | "careful" | 6; 21; 35 | 0.66; 0.66; 0.66 |
| | 0.8 - 0.9 | "tense" | 5; 20; 32 | 0.68; 0.64; 0.72 |
| | 0.8 - 0.9 | "greedy" | 5; 20; 32 | 0.66; 0.66; 0.66 |
| | 0.8 - 0.9 | boolean | 7; 22; 34 | 0.61; 0.72; 0.79 |
| | 0 - 0.2 | additive-weight | 3; 8; 16 | 0.69; 0.62; 0.71 |
| | 0 - 0.2 | "careful" | 3; 8; 14 | 0.61; 0.72; 0.72 |
| | 0 - 0.2 | "tense" | 3; 8; 16 | 0.71; 0.72; 0.71 |
| | 0 - 0.2 | "greedy" | 3; 10; 26 | 0.71; 0.68; 0.7 |
| | 0 - 0.2 | boolean | 5; 12; 22 | 0.61; 0.62; 0.59 |
| \( m_1 = 3m \) | 0.8 - 0.9 | additive-weight | 10; 30; 100 | 0.96; 0.96; 0.97 |
| | 0.8 - 0.9 | "careful" | 9; 28; 98 | 0.95; 0.98; 0.98 |
| | 0.8 - 0.9 | "tense" | 8; 26; 86 | 0.98; 0.95; 0.97 |
| | 0.8 - 0.9 | "greedy" | 10; 30; 100 | 0.99; 0.99; 0.99 |
| | 0.8 - 0.9 | boolean | 9; 30; 98 | 0.99; 0.99; 0.99 |
| | 0 - 0.2 | additive-weight | 10; 28; 96 | 0.98; 0.95; 0.97 |
| | 0 - 0.2 | "careful" | 8; 26; 92 | 0.96; 0.95; 0.98 |
| | 0 - 0.2 | "tense" | 7; 18; 72 | 0.96; 0.95; 0.98 |
| | 0 - 0.2 | "greedy" | 10; 28; 100 | 0.99; 0.99; 0.99 |
| | 0 - 0.2 | boolean | 10; 30; 100 | 1; 1; 1 |
| \( m_1 = 0.9m \) | 0.8 - 0.9 | additive-weight | 9; 26; 95 | 0.95; 0.95; 0.95 |
| | 0.8 - 0.9 | "careful" | 8; 24; 92 | 0.96; 0.95; 0.98 |
| | 0.8 - 0.9 | "tense" | 8; 24; 92 | 0.92; 0.95; 0.96 |
| | 0.8 - 0.9 | "greedy" | 10; 30; 100 | 1; 1; 1 |
| | 0.8 - 0.9 | boolean | 10; 30; 100 | 1; 1; 1 |
| | 0 - 0.2 | additive-weight | 9; 25; 94 | 0.92; 0.94; 0.91 |
| | 0 - 0.2 | "careful" | 9; 22; 92 | 0.9; 0.9; 0.86 |
| | 0 - 0.2 | "tense" | 8; 25; 94 | 0.87; 0.9; 0.9 |
| | 0 - 0.2 | "greedy" | 9; 29; 98 | 0.93; 0.98; 0.99 |
| | 0 - 0.2 | boolean | 9; 28; 97 | 0.99; 0.99; 0.99 |

In Table 1, the values for different dimensions of the characteristic space are listed through the ";" sign.
5. Discussion
The analysis of results of a computing experiment allows to draw the following conclusions:

- When using a statistically representative volume as a training and examination sample, the results of applying different approaches to the formation of a tuple of informative features do not significantly improve the quality of classification and, therefore, are unproductive.

- A similar effect is observed when the sample size is much higher than the representative threshold, but the requirements for computational resources (memory, speed) are increasing, therefore it is recommended to use the counter-buster method.

- The use of a set of proposed procedures for assessing the informative capacity for the formation of an informative tuple of characteristics makes it possible to obtain an acceptable classification result in the case of small sample use (statistically unrepresentative); the best way is the additive-weighted, "greedy" and boolean aggregations.

6. Conclusion
The offered technology of numerical nonparametric assessment of informational content of the measured signs of an object

The proposed technology of numerical non-parametric estimation of informational content of measured features of an object is recommended for practical use in conditions of weakly structured and / or statistically insufficient amount of data specific for the study of complex dynamic objects belonging to the class of open, complex systems.

The attribute cortege formed on the basis of estimates of information gives possibility to synthesize adequate mathematical models for solving the classification problems of objects based on the values of the measured features for monitoring and controlling the object in conditions of both a lack and an overabundance of measurement results characterizing the object.

References
[1] Taleb Nassim Nikolas 2017 Chernyj lebed'. Pod znakom nepredskazuemosti (Moscow: KOLibri, Azbuka-Attiku) Russian
[2] Kolesnikova S N 2009 Tomsk State university Bulletin Upravlenie, vychislitel'naja tehnika i informatika 1(6) p 48-54 Russian
[3] Visa Koivunen and Emad Mozafari 2016 50th Asilomar Conference on Signals, Systems and Computers Year (IEEE Conference Publications) pp 343–348
[4] Schmitt R., Fritz P and Lose J 2011 Int. J. Metrol. Qual. Eng. 2 (1) p 19–24
[5] Amos Tverski, Daniel Kaneman 2005 Vera v zakon malyh chisel (Kaneman D, Slovik P, Tverski A. Prinjatie reshenij v neopredeljonnosti: pravila i predubezhdienija) (Harkov: Gumanitarnyj centr Russian) p 40-41
[6] Juan Ramon Oreja-Rodrigues, Vanessa Yanes-Estévez, 2010 Proc. Environmental scanning: Dynamism with rack and stack from Rasch model (Management Decision) 48 (2) p 260-276
[7] GMDH – Examples of Applications Group Method of Data Handling (GMDH) 2010 available at: http://www.gmdh.net/GMDH_exa.htm
[8] Artemenko M V, Podval'nyj E S, Starcev E A 2016 Biomedicine radioelectronics (Biomedicinskaja radioelektronika) 9 p 38-44
[9] Richard S. Sutton, Andrew G 2016 Barto Reinforcement Learning: An Introduction (MIT Press)