Artificial Neural Networks as a Means of Restoring Passes in the Initial Data Array

A V Averchenkov¹, A A Kuzmenko¹, O V Stashkova¹
¹Bryansk State Technical University, Bryansk region, Bryansk 241035, Russia

E-mail: mahar@mail.ru

Abstract. The article presents an algorithm for restoration of the original data table using GRNN artificial neural network and the results of the algorithm in testing and empirical data. The given article also deals with calculations of the relative error for different data types with different percentages of passes. The quality of the analyzed data resulting from the passive experiment, as well as the reliability of the analysis results depends on one of the most important factors: the presence of these missing values. The distortion of the original data or completeness may distort the result in the general modeling process. Gaps in the original data table may be associated with a complete lack of data (raw data incompleteness) and contradictions arising from the data. And this kind of problem can occur not only with the values of a single attribute, but also with the values of a certain set of attributes especially in those cases when it comes to the large dimension of the factor space.

1. Formulation of the problem
A data table contains the results of test measurements of the parameters of one type of a certain process, wherein Par1..Par61 – parameters that are input factors (X1-X61). Such a table is a multidimensional (n-dimensional) sample. Each column of this table is the one-dimensional sample of a random variable X, derived from the 5305 independent experiments. This table certainly includes distorted values, and omissions, as expert’s work does not exclude subjectivity and/or trivial mistakes. It is necessary to explore the possibility for restoration of the original data table using an artificial neural network.

1.1. Analysis of the methods to solve the problem
It can be argued that the theory of recovery of missing data is constantly evolving and new or modified algorithms appear. The most common methods of handling incomplete information in data tables are [1, 4-6]:
- withdrawal of incomplete rows from a table or their replacement;
- filling of gaps with average values without selection with the (biased) selection;
- the method of the closest neighbors;
- regression method;
- the method of cluster analysis;
- maximum plausibility method and EM algorithm;
- ZET algorithm;
- Zet Braid algorithm;
2. Theoretical part
Analyzing the existing methods of recovery admissions it was decided to arrange the recovery of the missing values using artificial neural network (ANN) [7, 11-13, 15-19]. For this feature of neural networks was used the network which is able to find the relationship between the data. In other words, you can restore a single parameter if we know a few other examples of this training. In this work due to solving the problem of data recovery permits it was decided to use the generalized regression neural (GRNN) network which contains the following advantages:
- the ability to model nonlinear relationships between input and output parameters;
- network architecture is fixed and does not need a definition;
- learning time of network is much less than other INS.

Realized GRNN network has radially a base layer with the number of neurons equal to or less than the number of elements of the training set. ANN is represented by the following layers: a first middle layer network composed of radial elements; a second intermediate layer (line) containing elements that help to evaluate a weighted average; hidden layer generalized regression network, a block diagram is shown in figure 1.

![Figure 1. Hidden layer GRNN-network.](image)

The network establishes a first layer weight equal to $p$, wherein the offset is equal to 0.8326 divided by the distribution. The greater the spread the more smooth approximation function. Before you create a network there is a normalization of the input and target data. This is necessary for better network learning. Let us study the data recovery algorithm (figure 2) using artificial neural networks:
- checking of the initial table of statistics on the presence of gaps. In case of their presence we proceed to the second step of the algorithm;
- determination of rows and columns, in which there are missing values, data separation. We obtain complete data table, and a list of rows and columns with gaps;
- realizing of the process of machine learning network with complete data, that is the data in which there are no missing values;
- simulation of the work of the network on the samples which have missing values;
- obtaining a table of reduced values. Replace the empty values in the original table with the reconstructed values.

However, this method has negative sides: the loss of one of the training examples from the N possible. That is the total value of the training samples will be reduced by the number of examples with missing data. This can badly affect the result of the training network when training samples initially a small amount. The second disadvantage of this solution is that training example is deleted regardless of the number of missed parameters therein: It may happen so that one parameter of 20 may be passed or even 10 parameters.
3. Practical result
This algorithm has been implemented by means of an interactive environment Matlab program [2, 7-9, 14,]. The software product has been tested on a set of data presented in table 1. The work was carried out with four types of data.

Table 1. Types of the initial data of the test sample.

| Name of samples | Range   | Description                        |
|-----------------|---------|------------------------------------|
| Par_real        | 0..3,5  | The range contains values of the real type. |
| Par_500         | 0..500  | The range contains integer values with a large range of values. |
| Par_10          | 0..10   | The range contains integer values with a small spread of values. |
| Par_1_2         | 1; 2    | The range contains an integer binary value. |

The testing revealed that the INS does not work with quality indicators. In this case they should be converted into numerical values. This can be realized by association of a number. Also the lack of work INS is incorrect handling of values less than zero. Negative values ANN perceives as the absolute value of the initial value which distorts the results of the experiment.

Recovery data was investigated on the set with a different percentage of missing values. Tables 2, 3 present evaluation of samples with different percentages passes of various parameter types.
Table 2. Estimates of samples with 5% pass content.

| Passes                | Par_real | Par_500 | Par_10 | Par_1_2 |
|-----------------------|----------|---------|--------|---------|
| Dispersion            | 0.361409 | 20552,59| 6,983383| 0.249574|
| Standard deviation    | 0.481966 | 125,0314| 2,3164 | 0.498648|
| Average               | 0.816566 | 247,112 | 5,8412 | 1,474   |
| The confidence interval of the average value | 0.030238 | 7,844416 | 0.14533 | 0.031285|
| Absolute accuracy     | 0.008335 | 0.746693| 0.500457| 0.006965|
| Relative accuracy     | 1.02%    | 0.30%   | 8.57%  | 0.47%   |

Table 3. Estimates of samples with 75% pass content.

| Passes                | Par_real | Par_500 | Par_10 | Par_1_2 |
|-----------------------|----------|---------|--------|---------|
| Dispersion            | 0.361409 | 20552,59| 6,983383| 1,474   |
| Standard deviation    | 0.493624 | 124,2855| 2,33894 | 0.494422|
| Average               | 0.813865 | 245,982 | 5,835 | 1,4595   |
| The confidence interval of the average value | 0.03097 | 7,797619 | 0.146744 | 0.03102|
| Absolute accuracy     | 0.011035 | 0.383307| 0.495457| 0.021465|
| Relative accuracy     | 1.36%    | 0.16%   | 8.49%  | 1.47%   |

Table 4 contains a summary of the relative error values investigated depending on the percentage of gaps and the type of data to be restored.

Table 4. Relative accuracy for the investigated types with various percentage of passes (%).

| % of passes | Par_real | Par_500 | Par_10 | Par_1_2 |
|-------------|----------|---------|--------|---------|
| 5%          | 1,02     | 0,30    | 0,47   | 8,57    |
| 20%         | 0,05     | 0,09    | 0,44   | 8,69    |
| 50%         | 0,85     | 0,51    | 0,71   | 7,86    |
| 75%         | 1,36     | 0,16    | 1,47   | 8,49    |

Studies have shown that when recovering missed values, data with a large spread of values was most accurately recovered, the worst result was obtained when working with binary data. Regarding the number of recoverable values, the most accurate results are obtained with 20% of the missing data. When testing the above algorithm for recovering gaps in test data, the relative error of the average for each type of data did not exceed 9%.

The obtained results were tested on real process data in the form of a 61-column matrix (parameter) 5305 of rows (the number of experiments) obtained during the passive experiment (table 5). At the first stage, the values of the parameters were checked for rough errors (misses) [19, 20].

Sources of misses are often mistakes made by the operator in the measurement. The most characteristic of them are: incorrect reading on the scale of the measuring device, incorrect recording of the result of observation (misprision), incorrect recording of the values of individual measures of the used set, etc., errors in the operations with instruments if they are repeated during measurements. The causes of gross errors may be sudden or short-term changes in measurement conditions or unnoticed malfunctions in the equipment.
Table 5. Initial process data table.

| Par1 | Par 2 | ... | Par5 | ... | Par38 | ... | Par57 | ... | Par58 | ... | Par61 |
|------|-------|-----|------|-----|-------|-----|-------|-----|-------|-----|-------|
| 1,00 | 108,11| 14,29| ...  | 113,70| ...  | 0,09| ...   | 1748,67| 202,67| ...   | 94,00|
| 2,00 | 113,19| 13,40| ...  | 113,70| ...  | 0,12| ...   | 1746,50| 201,17| ...   | 97,33|
| 3,00 | 108,91| 13,01| ...  | 112,30| ...  | 0,10| ...   | 1748,83| 201,00| ...   | 91,50|
| 4,00 | 111,26| 17,18| ...  | 118,00| ...  | 0,10| ...   | 1748,83| 209,00| ...   | 104,0|
| 5,00 | 113,24| 15,60| ...  | 115,10| ...  | 0,11| ...   | 1749,00| 207,50| ...   | 102,0|
| 6,00 | 115,70| 13,77| ...  | 120,80| ...  | 0,10| ...   | 1747,33| 205,83| ...   | 99,50|
| 7,00 | 115,93| 13,77| ...  | 119,40| ...  | 0,11| ...   | 1746,50| 206,50| ...   | 100,67|
| 8,00 | 114,18| 15,20| ...  | 115,10| ...  | 0,10| ...   | 1746,33| 206,50| ...   | 100,83|
| 9,00 | 111,28| 13,40| ...  | 99,50 | ...  | 0,09| ...   | 1748,00| 213,00| ...   | 110,8|
| 10,00| 109,84| 17,20| ...  | 125,10| ...  | 0,10| ...   | 1747,60| 204,60| ...   | 97,20|
| ...  |      |      | ...  |      | ...  | ... | ...   | ...  | ...   | ... | ...
| 5302,0| 98,00| 38,90| ...  | 141,30| ...  | 0,11| ...   | 1749,40| 251,60| ...   | 87,40|
| 5303,0| 90,50| 47,50| ...  | 105,50| ...  | 0,13| ...   | 1730,67| 244,00| ...   | 81,17|
| 5304,0| 96,70| 42,30| ...  | 118,50| ...  | 0,13| ...   | 1748,50| 250,17| ...   | 85,67|
| 5305,0| 93,90| 42,90| ...  | 125,61| ...  | 0,11| ...   | 1748,33| 245,17| ...   | 81,50|

Suppose that the result of the observation X does not contain a gross error, i.e. is one of the values of the measured quantity. Using certain statistical criteria, it is possible to attempt to refute the hypothesis put forward. If this is possible then the result of the observations should be regarded as containing a gross error and must be eliminated or in any way eliminated.

There are a number of criteria that allow to exclude gross blunders [12]. These criteria include in particular the criteria of "three sigma", Grebbs (Smirnov), Sharlier, Shauven, Dickson, Romanovsky, etc. These criteria are based on statistical estimates of the distribution parameters, since in most cases the actual values of the distribution parameters are unknown.

In this paper when preparing the initial data for the training of ANN were excluded those values that do not correspond to the normal. In practice it is considered that if a rule of three sigmas is satisfied for any variable random then this variable random has a normal distribution: the absolute value of its deviation from the mathematical expectation does not exceed a tripled mean square deviation [12]. In fact, the results with coarse errors include either those that clearly do not correspond to the expected result of measurements or not pronounced extreme values the belonging of which to this array of results has a very low probability. All values that do not meet the «three sigma» criterion have been removed from the table, i.e. artificially created additional gaps in the source data table. As a result of the analysis of the original data table for «gross blunders» it was revealed that the total number of empty cells was 2.15%.

Table 6 presents the summary data on the relative error of the studied values.
Table 6. Parameter estimation.

| Parameter                                      | Par1     | Par2     | Par5     | Par38    | Par39    | Par57    | Par58    | Par61    |
|------------------------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|
| Average of the parameter in the source table   | 110,948  | 19,099   | 117,356  | 0,114    | 0,217    | 1754,264 | 214,990  | 99,764   |
| Average of the parameter, after recovery of the passes | 111,448  | 18,550   | 117,167  | 0,114    | 0,217    | 1753,920 | 214,867  | 99,870   |
| The relative error of the parameter (%)        | 0,449    | 2,963    | 0,161    | 0,001    | 0,005    | 0,020    | 0,057    | 0,107    |

The relative error in the recovery of experimental data of the technological process of metallurgical production was 1.31%.

4. Conclusion
The article proposes a method of filling the gaps in the table of input data obtained as a result of a passive experiment based on artificial neural networks in particular the GRNN network. The comparative analysis of the ANN operation with data of various types is performed at the same time the relative error of the investigated values is estimated depending on the percent of passes and the data type. An example of realization of approach is considered on data of technological process thus with their preliminary analysis on «gross blunders». It is necessary to mark the sufficient exactness of INS work: at a sample size of 323605 values the index of relative error made no more than 2%. Based on the results of the study it can be argued that ANNs can be used for adequate data recovery.

References
[1] Abramenkova I V and Kruglov V V 2005 Methods of recovering passes in data sets *Software products and systems* 2 pp 18-22
[2] Averchenkov A V and Stashkova O V 2017 Simulation of artificial neural networks for solving problems of factor analysis *Innovations in industry, management and education: a collection of scientific papers of the international scientific-practical conference* (Bryansk: FSBEI HE "BSTU") pp 117-120
[3] BaseGroup *Data analysis technologies* https://basegroup.ru
[4] Efimov A S 2010 Solution of the problem of clustering by the method of concrete training with incomplete statistical data *Bulletin of the Nizhny Novgorod University N I Lobachevsky* 1 pp 220-225
[5] Martyshenko S N 2013 Methods for recovering omissions in data presented in various measuring scales *Territory of new opportunities* 4(22) pp 242-255
[6] Metrology, standardization and certification http://studme.org/11211219/tovarovedenie/grubyepogreshnostipromah
[7] *MATLAB Documentation* http://www.mathworks.com/help
[8] MathWorks *MathWorks product documentation* http://www.mathworks.com
[9] Matlab and Toolboxes *Materials on MATLAB & Toolboxes products* http://matlab.exponenta.ru
[10] Rogachev A F and Shubnov M G 2013 Construction of neural network models for predicting time series yields based on autocorrelation functions *Modern problems of science and education* 5 p 450
[11] Snityuk V P 2006 Evolutionary method for recovering gaps in data Collected Works of the VI-th International Conference «Intellectual information analysis» (Kyiv) pp 262-271

[12] 2001 Stochastic check for control of electronic wares quality Trans. of 10th International Symposium on Applied Stochastic Models and Data Analysis Univ. de Techn. de Compiègne (France) vol 1 pp 387-390

[13] Stashkova O V 2017 Measure of information uncertainty as a condition for the reduction of sampling parameters Mathematical modeling in education, science and production: Coll. of abstracts of reports of the X International Conference September 28-30, 2017 (Tiraspol) pp 43-45

[14] Stashkova O V 2018 Development of a complex of programs for modeling a multifactor object New information technologies in education and the agricultural sector of the economy: A collection of materials of the I International Scientific and Practical Conference (Bryansk: Publishing house Bryansk State Agrarian University) pp 270-273

[15] Neural Network Software, About NeuroSolutions Available at: http://www.neuroproject.ru/aboutproduct.php (circulation date: 04/19/2018)

[16] Burakov V M 2013 Neural networks and neurocontrollers: tutorial (SPb.: GUAP) p 284

[17] Neural networks STATISTICA Neural Networks: Methodology and Technologies of Modern Data Analysis Edited by V P Borovikov 2nd ed. (M.: Hotline – Telecom) p 288

[18] Hornik K ana Stinchcombe M and White H 1989 Multilayer feedforward networks are universal approximators Neural Network vol 2 p 366

[19] Khaikin S 2006 Neural Networks: Full Course = Neural Networks: A Comprehensive Foundation 2nd edition (M.: Williams)

[20] Kozak L Y 2011 Processing of primary data on indicators of the technological process of smelting high-carbon steel Materials of the International Scientific-Practical Conference (Rybnitsa, Tiraspol: Publishing house Pridnestrovie University)