Choosing a statistical method for predicting a quantitative indicator

R G Vlasov¹ and Yu S Korobov¹,²
¹Department "Welding Production Technology" of the Ural Federal University named after B. N. Yeltsin, Yekaterinburg, 620002, Russia
²Institute of Metal Physics named after A. N. Mikheev, Ural Branch of the Russian Academy of Sciences, Yekaterinburg, 620002, Russia

E-mail: novik904@yandex.ru

Abstract. The aim of the study is to find a machine learning method that allows achieving the maximum accuracy of the forecast when training on a specific data set. The article provides a comparative overview of strategies for choosing the optimal machine learning method. The process of training sample preparation is considered, the necessity of its modification is justified. The search for the optimal machine learning method was carried out in the statistical computing environment "R" using the package "CARET". The graphical results of training 96 algorithms that implement various statistical methods are presented, and the choice of the best one is justified on the basis of the proposed methodology.

1. Introduction

The main tool for data analysis is machine learning, and the number of methods used in modern machine learning is in the hundreds. In the process of problem solving of choosing the optimal machine learning method and the algorithm that implements it, we analyzed the strategies used by various authors to select the appropriate method.

- The most popular strategy is to analyze publications on the topic of interest and choose from several popular methods by empirically comparing their effectiveness in relation to the data under study [1-8].
- Selecting several of the most popular methods in a particular class and making an ensemble of them [9].
- The choice of the method depending on the type of data studied, in accordance with the set of some rules [10-12].
- The use of the most popular method without justifying the reasons for its choice [13-18].

From the point of view of the authors of the article, strategies 1, 3 and 4 cannot be considered satisfactory, since most of the machine learning methods (possibly more effective) are not even considered when implementing such a search.
2. Initial data

As the initial data for the mathematical model construction, data describing the process of testing technically complex products of several types in the assembly shop of a machine-building enterprise are used. Table 1 shows the dependence between the test duration of the product (response \((y)\)) and the number of other products already in the test process at the start time of the product's test (vector of input parameters \((x_1...x_{10})\)). In total, there are 50 observations in the sample.

| № | Vector of input parameters \((x_1...x_{10})\) | Response \((y)\) |
|---|---|---|
| 1 | 2 1 0 2 1 1 0 0 2 0 | 40 |
| 2 | 0 0 0 2 0 0 1 2 3 | 54 |
| 3 | 1 0 0 2 0 0 1 2 3 | 47 |
| 4 | 3 0 0 2 0 0 1 2 3 | 52 |

The exploratory analysis of the training sample was carried out according to the protocol [19], which consists in a step-by-step verification of the statistical criteria of the original sample. It was found that some variables have near-zero variance (they have little variability—they contain many values of "0" and only a few values of "1"). Such variables can have a negative impact on the convergence of those statistical methods that calculate the covariance matrix in the process. In order to study the effect of zero variance dispersion on the convergence of various methods, three training samples were formed according to the following rules:

1. Sample "C" - the original set (Table 1. Initial data for modeling).  
2. Sample "A" (Table 2. Sample "A"): a vector consisting of a repeating sequence of numbers from 0.001 to 0.009 is added to each column;  
3. Selection "B" (Table 3. Selection "B"): a vector consisting of a repeating sequence of numbers from 0.01 to 0.09 is added to each column.

| № | Vector of input parameters \((x_1...x_{10})\) | Response \((y)\) |
|---|---|---|
| 1 | 2.001 1.001 0.001 2.001 1.001 1.001 0.001 0.001 2.001 0.001 | 40 |
| 2 | 0.002 0.002 0.002 0.002 2.002 0.002 0.002 1.002 2.002 3.002 | 54 |
| 3 | 1.003 0.003 0.003 0.003 2.003 0.003 0.003 1.003 2.003 3.003 | 47 |
| 4 | 3.004 0.004 0.004 0.004 2.004 0.004 0.004 1.004 2.004 3.004 | 52 |
3. Research method
The statistical computing language "R" and the software package "CARET" were used as a tool for constructing mathematical models [20]. The "CARET" package allows the automatically finding of the optimal values of hyperparameters by using cross-validation to test the quality of each model. The developers have provided the default values of hyperparameters for each algorithm, so using this package gives a significant gain in time, since it eliminates the need to configure a large number of hyperparameters of each model iteratively, in manual mode. At the time of writing this paper, the package included over 200 statistical learning algorithms from various packages. We have selected 98 algorithms that represent various machine learning methods and are most often used for solving regression problems.

4. Results and discussion
Using three training data sets, 98 algorithms were trained, and the corresponding mathematical models were constructed. The models were evaluated using the metrics $R^2$ (R-squared – coefficient of determination) and $MSE$ (Mean square error – mean square error of the forecast). $R^2$ can be interpreted as a measure of the model’s compliance with the data, and $MSE$ as an offset-the deviation of the predicted values from the training sample values.

$MSE$ calculated as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$  \hspace{1cm} (1)

where: $\hat{y}_i$ – predicted value of the quantitative response; 
$y_i$ – quantitative response from the training sample; 
n – number of observations.

$R^2$ can be calculated as:

$$R^2 = (r_{yx})^2$$  \hspace{1cm} (2)

where $r_{yx}$ – the correlation coefficient, calculated as:

$$r_{yx} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n \sigma_x \sigma_y}$$  \hspace{1cm} (3)

where $\sigma_y$ and $\sigma_x$ mean square deviations for y and x, respectively:

$$\sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}; \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$  \hspace{1cm} (4)

It is important to note: in order to save resources, $MSE$ (1) and $R^2$ (2) were calculated using only the training sample, without checking on the test sample.
Based on the results of training 96 models, diagrams in the $MSE$ and $R^2$ axes are calculated, which are shown in Figures 1, 2 and 3.

The main goal of the study is to find a method that will allow us to obtain a model with minimal bias and most fully describe the dependence between $x$ and $y$ (match the data). That is: the model must have a maximum $R^2$ and a minimum $MSE$, and therefore, its coordinates will lie in the upper-left corner of the diagram. Also, the analysis of Fig. 1-3 allows us to draw conclusions about how stable the algorithms representing different machine learning methods work when using training samples with different forms of representation of input variables. Figure 1 shows a rectangular area where models with $MSE = 0$ are located. These models are obtained using algorithms known as "Nuclear Smoothing Methods". The value of $MSE = 0$ indicates the inadequate operation of these algorithms, confirms the assumption that variables with near-zero variance have a negative impact on the convergence of some methods, and confirms the need to modify the input variables. Most algorithms have built predictive models (oval area). Dashed lines show models with the minimum $MSE$, and an arrow marks the model with the maximum value of $R^2$.

![Figure 1. Results of training algorithms using the sample "C".](image)

In figure 2: it is obvious that as a result of training using the "A" sample, most of the algorithms have built predictive models: there are no models with $MSE = 0$. The exception is two algorithms that have $R^2 = 0$, and these are again "Nuclear Smoothing Methods".
There is a group of models with average accuracy (oval area), but in general, the results are worse than in figure 1, since there are much more models with a relatively large $MSE$. The model with the maximum value of $R^2$ is marked with an arrow.

Figure 3 clearly shows the winner-the model in the upper-left corner of the diagram, marked with an arrow. This model is obtained by a neural network-a multilayer perceptron with the topology $10 – 4 – 5$.
– 1 (10 neurons in the input layer, 4 and 5 in the first and second hidden layers, respectively, and 1 output neuron).

The coordinates of the best model $R^2 = 0.82$ and $MSE = 6.98$ indicate that the algorithm has identified and described quite well the dependence between the vector of input parameters (x1...x10) and the response Y. At the same time, it is worth paying attention to the model whose coordinates are slightly lower (marked with an arrow): this is a Bayesian neural network with topology 10 – 10 – 1, which can potentially get a model with similar indicators.

In our opinion, it will be preferable in the case of a more complex sample consisting of hundreds of observations and the presence of noise or outliers in the sample. The group of models with average accuracy values is somewhat more blurred than in figure 1 and figure 2. There is also a group of models that has the value $R^2 = 0$, which indicates a failure in the process of building models by algorithms. All these algorithms belong to different methods and you can determine the causes of the failure only by studying the principles of their operation in detail. At the stage of finding the optimal algorithm, these methods can simply be excluded from the list of tested ones.

5. Conclusion
The proposed method of evaluating the ability of machine learning algorithms to describe the analyzed data allows you to make an informed choice in the direction of the optimal algorithm, contributes to obtaining a more accurate model and significantly saves the time required to achieve the expected result. Also, the method allows you to assess the need for data pre-processing and the effectiveness of the processing itself. This distinguishes it from the strategies described above. The only significant drawback of the method is the resource intensity.

References
[1] Aizpurina J I and others 2017 IEEE Conference on Electrical Insulation and Dielectric Phenomenon (CEIDP). 24(4) 153-6. doi: 10.1109/CEIDP.2017.8257475
[2] Ballester P J 2020 Drug discovery today: technologies. 32 81-7. doi:10.1016/j.ddtec.2020.09.001
[3] Guolang S, Mufan L and others 2020 Hindawi: Mathematical Problems in Engineering 1-13. doi:10.1155/2020/7842847
[4] Khaledian Y, Miller B and others 2020 Applied Mathematical Modelling. 81 401-18. doi:10.1016/j.apm.2019.12.01
[5] Tarasov O F and Turlakova S S 2018 Industrial economics. 3(83) 57-75. doi: 10.15407/econindustry2018.03.057
[6] Staroverova K Y and Bure V M 2017 Bulletin of the Saint Petersburg University. Applied mathematics. Computer science. Management processes. 13(1) 51-60. doi: 10.21638/11701/spba10.2017
[7] Dikij D I 2020 Information and control systems. 4 50-60. doi:10.31799/1684-8853-2020-4-50-60
[8] Mayfeld H J, Smith C and others. 2020 Environmental Modelling and Software. 131 1-10. doi:10.1016/j.envsoft.2020.104741
[9] Shibajkin S D, Nikulin V V and Abbakumov A A 2020 Bulletin of the Astrakhan State Technical University. Series: Management, computer engineering and Computer Science. 1 29-40. doi: 10.24143/2072-9502-2020-1-29-40
[10] Doronina L S and Kaskinova D I 2017 Proceedings of the Orenburg State Agrarian University. 7(7) 49-52
[11] Neminushhaya L A, Skotnikova T A and others. 2015 Bulletin of the Technological University. 18(2) 377-82
[12] Narkevich A N and Vinogradov K A 2019 Social aspects of public health 65(4). URL: http://vestnik.mednet.ru/content/view/1092/30/lang.ru. doi:10.21045/2071-5021-2019-65-4-9
[13] Ablaev M F and Salixova N M 2020 Scientific Notes of the Kazan University. Series of Physical and Mathematical Sciences. 162(3) 41-58. doi: 10.26907/2541-7746.2020.3.241-258
[14] Xal`yasmaa A I 2020. *Bulletin of the Irkutsk State Technical University*. **24**(5) 1093–104
doi:10.21285/1814-3520-2020-5-1093-1104

[15] Bure N A, Grebennikova N L and Staroverova K 2018 *Bulletin of the Saint Petersburg University. Applied Mathematics*. **14**(4) 325-33

[16] Kadomcev M I and Razveeva I F 2020. *Engineering Bulletin of the Don. 2* (62) 12-8.

[17] Xajrullina A I, Madzhidov T I and others 2018. *Scientific Notes of the Kazan University. Series of natural sciences*. **160**(2) 200–13

[18] Sulimov P A 2016 *Bulletin of the Novosibirsk State University of Economics and Management*. **1** 285-306

[19] Alain F Zuur , Elena N Ieno and Chris S Elphick. 2010 *Methods in Ecology and Evolution*. **1**(1) 3–14. doi: 10.1111/j.2041-210X.2009.00001.x

[20] Site: The “caret” Package: https://topepo.github.io/caret/index.html