The use of aggregate classifiers in technical diagnostics, based on machine learning

V.N. Klyachkin[1], Yu.E. Kuvayskova[1], D.A. Zhukov[1]

[1]Ulyanovsk State Technical University, 432027, Ulyanovsk, Russia

Abstract

While solving the problems of technical diagnostics with machinery learning involvement, there is binary classification of an object state performed: the objects are subdivided into “good” and “bad” with the help of models, received as per learning samples. The quality of classification, which specifies the efficiency of machine learning, depends on several factors, such as: the scope of original sample, method of machine learning, method of dividing the sample into learning and validating parts, selection of value indicators, etc. Sometimes it is reasonable to use aggregate methods of classification, which are, in fact, the joined results of classification basic methods. To search the best aggregate method, one iterates over all possible basis sets.

Keywords: binary classification; “good” and “bad” state; aggregate method

1. Introduction

The object technical diagnostics is done to increase the reliability of the system, and it is often limited by the assessment of its serviceability [1,2]. The main aim is object state recognition. By recognition we mean an object state, typed as per one of the classes - diagnoses .As a rule, the problem solution is restricted by classifying the object as good, i.e. able to perform the desired functions, or bad. This is the task of binary classification. The selection of parameters, characterizing the system state, is important. The assessment of the system state is done during its operation, the information receiving is somehow difficult, to take a decision - different methods of recognition are used. The solution of technical diagnostics problems is also connected with the technical object state forecasting [3,4].

The recognition of the object technical state is usually done, based on the results of indirect indicators of the object operation under the conditions of available limited information. The known results of the system state assessment are used: for the selected values of monitored parameters the system is assessed as “good” or “bad” (serviceable or unserviceable). So, there are many objects (situations) with selected indicators, and many possible states of the system. There is acertain unknown dependence between the object operation indicators and its actual sate. The finite universe of pair - “set of indicators, state”- is known, i.e. the original data retrieving. It is required ore store the dependence, i.e. to build an algorithm, capable to generate a rather accurate response. Anyway there is a risk of getting false warning or missing the target. This is the task of machine learning, or learning from examples (with a tutor) [5,6].

To measure the accuracy of classification there is performance functional introduced, e.g. the mean error can be used: original sample is divided into learning one, with the help of which the algorithm of the needed dependence is identified , and validating one (test), with the help of which the mean error is assessed [7,8].

2. Methods of machine learning , used for binary classification

The methods of machine learning are widely used indifferent fields: speech recognition, medical diagnose, loan score and others. From the point of view of technical diagnostics, the machine learning is only binary classification task: as per the selected vector of object parameters, it is necessary to determine, which sate of the object is (“bad” or “good”).

For solving the problems of technical diagnose, the method of learning from the examples is used. Bayesian classifier is one of them, as well as K-Nearest Neighbors (KNN) method, neural network, logistic regression, discriminate mining, Support Vector machine method, decision trees, etc.

The problem of technical object classification is solved as follows: the object is found “good” *Y* = 1, if the probability model is $P\{Y=1\,|\,X\}>0{,}5$, and “bad” *Y* = 0 – if opposite. As threshold any number different from 0,5 can be used.

For example:

Using logistic regression, we assume that the probability of “good” object is equal to *Y* = 1:

$$P\{Y=1\,|\,X\}=f(z),$$
$$z=q_0+q_1x_1+...+q_nx_n, \qquad (1)$$

where $q_0,...,q_n$ is model parameters (1), *f*(*z*) is logistic function:

$$f(z)=\frac{1}{1+e^{-z}}. \qquad (2)$$

As variable *Y* takes one of pair values (0,1), the probability of “bad” state is equal to *Y*= 0:

$$P\{Y=0\,|\,X\}=1-f(z). \qquad (3)$$

So, the logistic regression is based on the following expression:

$$\log\frac{P\{Y=1\,|\,X\}}{P\{Y=0\,|\,X\}}=\frac{f(z)}{1-f(z)}=q_0+q_1x_1+...+q_nx_n. \qquad (4)$$

To identify parameters $q_0,..., q_n$, as a rule, maximum likelihood method is applied . This method is aim edat likelihood function maximization with the help of gradient descend method, Newton Raphson method and others.

With the time, while new data are received, the earlier found parameters can become out-dated. Toup-date them, different procedures can be used, e.g. those based on pseudogradient use [9].

The short coming of the logistic model is its sensitivity to factors correlation, that is why, the presence of strongly correlated input variables is unacceptable in the model.

The advantage of them odel is the possibility to take in to consideration the limitations of probability value, which cannot be out of frame 0 and 1, the possibility of conducting investigation and assessment of the factors, affecting the result.

While using another widely applied model – discriminant mining– to determine the object *m* class, linear discriminant functions are used:

$$
\begin{aligned}
&o_1(x) = q_0^1 + q_1^1 x_1 + ... + q_n^1 x_n,\\
&o_2(x) = q_0^2 + q_1^2 x_1 + ... + q_n^2 x_n,\\
&...\\
&o_m(x) = q_0^m + q_1^m x_1 + ... + q_n^m x_n,
\end{aligned}
\tag{5}
$$

Where *o*(*x*) is «counting», as per which this or that class is identified. In result, that class is chosen, which counting is the highest. The model parameters are assessed with the help of learning sample. In case of two classes, there sultcoincides with the result of linear regression.

The point is, that, in advance, one cannot say, which method, out of those mentioned above, will ensure the correct solution of the problem, that is why several methods or combination of methods are used. The decision is taken, based on the results of the performance functional for the validation set.

3. Aggregate methods

Aggregating methods (combined application of several methods) is of special interest, indeed, as this way compensates the disadvantages of one model with the help of the others, thus improving the forecasted accuracy. Now we will consider the follow in gset, let us say, of base 7 models [10,11]: neural network, logistic regression, discriminate mining, Bayesian classifier, Support Vector machine method, decision trees, bagging trees etc.

Let's make all possible combinations of base models, consisting of two, three… etc. models. In case of seven models taken, the total amount of all the combinations, starting from two and finishing with all seven models, will make: $C_7^2 + C_7^3 + C_7^4 + C_7^5 + C_7^6 + C_7^7 = 120$ models.

Let Y_j^m be the result of serviceability assessment of j-object. The result was determined with *m* base model, $j = 1,...,l$ and $m = 1,...,M,$ where *M* – the number of base models in combination. Now let us see the following ways of base models aggregating (joining).

a. *Aggregating ,based on mean value*

In this case

$$
Y_j^{AK\_mean} = \frac{\sum_{m=1}^{M} Y_j^m}{M}, \tag{6}
$$

where $Y_j^{AK\_mean}$ is the result of aggregate classifier based on mean value.

b. *Aggregation as per median value*

Firstly, we will grade the row, containing the results of the base models in combination Y_j^m. If the number of base models is odd:

$$
Y_j^{AK\_median} = Y_j^{\frac{M+1}{2}}, \tag{7}
$$

where $Y_j^{AK_median}$is the result of aggregate classifier as per median value.

c. *Aggregation as per voting*

This works as follows: if the majority of the models consider the object is "good", then the result of aggregate classifier is a mean value of the results of the models, voting for the serviceable class. In opposite case, the object is "bad" (unserviceable) (*Y*=0).

4. Diagnose quality assessment

The accuracy of classification is assessed with the help of performance functional, i.e. validations set classification mean error can be used.

When the results are presented in the terms of object “good’ or “bad” class probability, to assess the methods quality it is possible to find error dispersion σ^2, which indicates the deviation of forecasted value from actual ones:

$$\sigma^2 = \frac{1}{l}\sum_{r=1}^{l}(P(Y_r) - \hat{P}(X_r))^2, \tag{8}$$

where $P(Y_r)$ is actual probability of serviceability class of r object ($P(Y_r)$= 0, if the object is “bad” or $P(Y_r)$= 1 for the “good” object), $\hat{P}(X_r)$ is actual probability of serviceability class of r object, l is number of objects.

The performance functional mainly depends on the way of validation set constructing. If necessary, check the influence of the way of validation set constructing on the error dispersion.

In order to optimize the diagnostic of technical object functioning, the algorithm of serviceability forecasting is offered. This is the use of machine learning models combination and generating optimum decision on their basis.

The algorithm main stages:

1. Generating and preliminary processing of the original data, dividing them into learning and validation set.
2. Constructing the base classification model on a learning set.
3. Building all the possible models combination on the same learning set in addition to all base models with three mentioned above methods aggregation.
4. On the validation set, through all the constructed models, new (monitored) objects serviceability is forecasted.
5. For each model or model combination, the forecasted mean square root error is calculated, and the best model, providing the error minimum, is selected.

5. Results and discussion

The computational investigation was done based on the example of St. Petersburg sewage plant operation [12].The parameters of the water supply source and the dosage of chemical agent, used for purifying, were monitored. At least one parameter of potable water quality, found out beyond the acceptable limitations, was considered to be the system malfunction. Seven base methods of binary classification were used, and the best result was shown by the method of support vector machines (SVM); mean classification error was equal to 0,238. Mean value aggregation had an error equal to 0,196 (combination of SVM with discriminate mining and neural network). While using the selection of tangible value parameters through the method of stepwise regression, the results deteriorated a little, but even that, minimum mean error of aggregation was 0,207. The set of base classification changed: the logistic regression method was added to those three available.

The performance quality of classification is determined by the scope of the original sample , selected by machine learning method (one of base or aggregate), by the method of dividing the original sample into learning and validation one (either by random selection, or by taking a certain part of original sample for validation one; sometimes the procedure of sliding exam is reasonable, evidently, the scope of validation sample plays a certain role hereto), method of tangible value (e.g. stepwise regression) and some other factors.

To provide the efficiency of machine learning for the technical object diagnostic, it is necessary to work out the system in order to investigate the influence of these factors on the performance quality of classification with original sample, which could ensure the optimum approaches.

6. Conclusion

The given above investigation showed, that the methods of machine learning could be used for solving the problems of technical diagnostics, i. e. identifying the state of serviceability of the investigated object. Here some problems may arise. The problems are those connected with generating a rather big scope original sample, with dividing the sample into learning and validating, with assessment of this or that method efficiency , with possibility to use aggregate classifiers .

Acknowledgements

The investigation is done with financial support of RFFI (РФФИ), projects №16-48-732002 and №16-38-00211 mol_a.

References

[1] Birger IA. Technical diagnostics. M.: Mechanical engineering, 1978; 240 p.

[2] Zhukov DA, Klyachkin VN. The tasks of machine learning efficien cyprovision for technical objects diagnostics. Modern problems of radio aids design, industry and operation 2016; 1(10): 172–174.

[3] Klyachkin VN, Bubyr DS. Forecasting of technical object state based on piecewise linear regression. Radiotechnics 2014; 7: 137–140.

[4] Klyachkin VN, Kuvayskova YuE, Bubyr DS. Forecasting of technical object state with the use of time series system. Radiotechnics 2015; 6: 45–47.

[5] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005; 525 p.

[6] Merkov AB. Patterns recognition. Introduction to the metods of statistical learning. M.: Editorial URSU, 2011; 256 p.

[7] Klyachkin VN, Kuvayskova YuE, Alekseeva VA. Statistical methods of data mining. M. : Finance and Statistics, 2016; 240 p.

[8] Klyachkin VN, Karpunina I.N, Kuvayskova YuE, Khoreva FR. Machine learning method in technical diagnostics. Academic Bulletin of UCAS 2016; 8: 158–161.

[9] Krasheninnikov VR, Klyachkin VN, Shunina YuR. Updating of agggregate classifiers on the basis of pseudogradient procedure. Academic Bulletin of computer and information technologies 2016; 10(148): 36–40.

[10] Shunina YuR, Klyachkin VN. Forecasting of bank customers creditobility on the basis of machine learning methods and Markov chains. Software products and Systems 2016; 2: 105–112.

[11] Shunina YuR, Alekseeva VA, Klyachkin VN. Forecasting of bank customers creditability on the basis of machine learning method. Finances and Credit 2015; 27(651): 2–12.

[12] Kuvayskova YuE, Bulyzhev YeM, Klyachkin VN, Bubyr DS. Forecasting of water supply source state in order to ensure water quality. Reference manual. Engineering pamphlet with attachment 2016; 5: 37–42.