

# Selection of aggregated classifiers for the prediction of the state of technical objects

D A Zhukov<sup>1</sup>, V N Klyachkin<sup>1</sup>, V R Krashennnikov<sup>1</sup> and Yu E Kuvayskova<sup>1</sup>

<sup>1</sup>Ulyanovsk State Technical University, Severny Venets street, 32, Ulyanovsk, Russia, 432027

e-mail: zh.dimka17@mail.ru, v\_kl@mail.ru

**Abstract.** The basic data in the problem of the prediction of technical object's state of health based on the known indicators of its operation are the known results of the object state estimation by information about previous service. The problem may be solved using the machine learning methods, it reduces to binary classification of states of the object. The research was conducted in the Matlab environment, ten various basic methods of machine learning were used: naive Bayes classifier, neural networks, bagging of decision trees and others. In order to improve quality of healthy state identification, it has been suggested that aggregated methods combining several basic classifiers should be used. This paper addresses the issue of selection of the best aggregated classifier. The effectiveness of such approach has been confirmed by numerous tests of real-world objects.

## 1. Introduction

It is possible to forecast the state of the technical object using various methods. The realistic simulation using the time-series system is the most commonly used approach [1-4]. However, as often as not forecasting comes to the object state division in the target horizon in operating ones, i.e. capable of fulfilling intended functions, or faulty ones. Still and all, the diagnostics is carried out according to the object operation and the measurement of indirect values of its functioning.

For example, the engine performance is diagnosed by reference to the fuel consumption rate, the gas temperature, the noise and vibration level, the exhaust gas composition, the clearance between the cylinder and the piston, the clearance between crankshaft necks and bearings and some other indicators [5]. Therein, there is a false alarm risk (when the operating object will be considered as the faulty one) or vice versa when the faulty object is considered as the operating one will be skipped.

Basic data are a priori information about the state of the object according to the results of the previous exploitation: upon the given values of controlled indicators the technical system is operating one or the faulty one. It is assumed that there is some unknown dependence between indicators of the object functioning and its states. Based on basic data it is necessary to restore this dependence, i.e. to plot an algorithm, capable of providing a fairly valid answer about the state of the object for the given set of indicators of its functioning. It is a task for the computer-aided learning or the learning from examples (with a tutor). Binary classification, i.e. the object state division in two categories [6-8], is a special case of this task.

To assess the quality of the plotted algorithm in the context of the opportunity to forecast, the original sample is divided in two disjoint subsets. The first subset is the learning sample itself for handling the task of learning (which, usually, comes to assessing parameters of the model of the appropriate algorithm). The second subset is the control (or test) sample which is not used for learning. This part of the sample is used estimate the forecast error which characterizes the quality of learning. When using the cross-validation, the sample is divided in  $N$  parts (in practice, usually,  $N = 5$  or  $N = 10$ ). In this case,  $N - 1$  parts are used for learning and the rest for control. All possible options are sorted out successively.

Methods of the computer-aided learning are actively used in all kinds of activities. Many different approaches to the classification are used. For example, classical statistical methods (Bayesian classifiers, the discriminative analysis, the logistic regression), methods specially focused on the computer-aided learning (the support vector machine, neural networks), compositional methods (bagging, boosting) and etc. The question at issue that it is impossible to determine which method from selected ones will provide the best solution of the task. That is why many different methods or their combinations are usually used. Decision to apply is made based on findings of the research of the quality functional for the control sample. In works [9-10], the aggregate approach, applying of the combination of several classification methods, is suggested to improve the forecasting quality. These results were certified by the experiment and for technical diagnostics tasks as well [11-13].

The purpose of this study is to plot selection algorithms of the best aggregated classifier.

## 2. Using of basic classifiers

The most widely known indicator that can be used for the quality assessment of the binary classification is the proportion of correct answers in the control sample,

$$Accuracy = \frac{Q}{N},$$

where  $Q$  is the number of correctly classified objects from the control sample and  $N$  is the overall control sample size. The opposite characteristic which is the proportion (or the percentage) of errors in the control sample is used more often.

Sometimes the error dispersion (the mean square deviations of the operating state true probability in the  $r$ -test  $P(Y_r)$  from its forecasted value  $\hat{P}(X_r)$ ) is used to assess the quality of the classification:

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

If classes are unbalanced (when there are much more operating states than faulty ones), the proportion of errors cannot be used for the reliable quality assessment of the classification [15-16]. Accuracy and completeness are far more informative

$$P = \frac{tp}{tp + fp},$$

$$R = \frac{tp}{tp + fn},$$

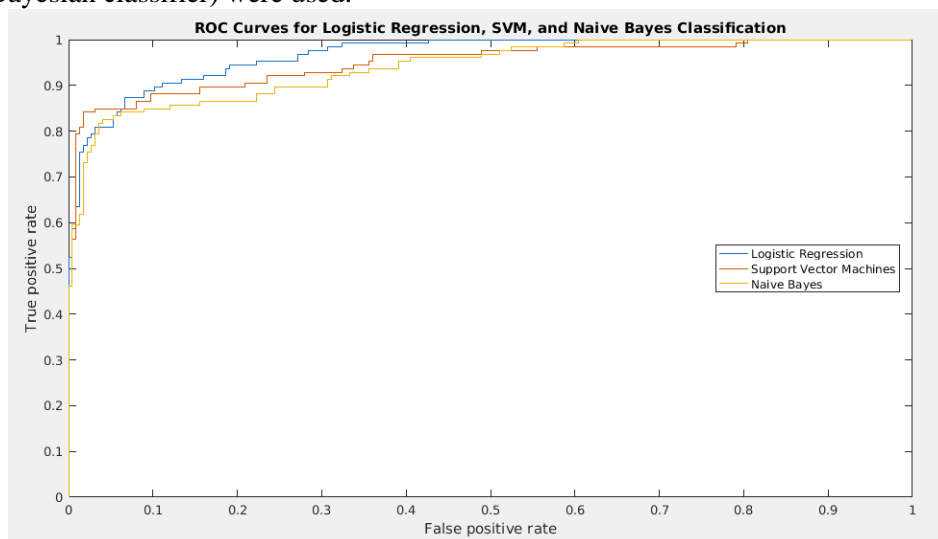
where  $tp$  is the number of properly classified operating states,  $fp$  is the number of misclassified operating states,  $fn$  is the number of misclassified faulty states. Based on these two indicators the uniform criterion can be formed.

$$F = \frac{2PR}{P + R},$$

– it is called the harmonic average of the accuracy and the completeness ( $F$ -measure): the closer is the value of  $F$  to one, the higher is the quality of the classification.

Area under the ROC-curve (receiver operating characteristics): AUC (area under the curve) can also be selected as the quality functional [16-20]. ROC-curve will be formed, if values  $fp(c)$  are taken at the x-axis and values of  $tp(c)$  are taken at the y-axis, where  $c$  is the threshold. Area under the ROC-curve gives an opportunity to assess the model in general without being related to the certain threshold. Criterion AUC-ROC is resistant to the influence of unbalanced classes. It can be interpreted

as the likelihood that the probability value of the randomly selected object from the class 1 will be closer to 1 in comparison with the randomly selected object from the class 0. Such curves are shown in Fig. 1. They are plotted in the Matlab system for the diagnostics example considered below. In this case three methods of the binary classification: the logistic regression, the support vector machine and the naïve Bayesian classifier) were used.



**Figure 1.** ROC- curves for three classification methods.

As an illustration of the numerical study we considered the water treatment system. We had the results of 348 tests upon eight quality indicators of the drinking water treatment. The system was faulty in 47 cases (when even one water quality indicator was beyond the limits). Whereas the division of basic data in the learning sample and the control sample was carried out randomly, we repeated tests 50 times.

We used the Matlab-package for tests. In Table 1 there are averaged values of the  $F$ -criterion and the area under the ROC-curve AUC for those five methods of the computer-aided learning where these values were maximum. Estimates suggest that the correlation between these two indicators was non-significant at the significance level 0.05. If the  $F$ -criterion is the same for selected classifiers, AUC values can be used for selection of the best classification method.

It is apparent that the decision tree bagging showed the best results in the considered example.  $F$ -criterion discrepancy between the best and the worst (0.801 for the RUSBoost method) results was 8.7%, AUC – 21.5%.

**Table 1.** Quality measures of various classification methods.

	$F$ -criterion	AUC
Neural network	0.836	0.822
Decision tree bagging	0.871	0.893
Gradient boosting	0.860	0.862
AdaBoost	0.852	0.854
Logistic regression	0.844	0.870

### 3. Aggregated classifiers

The aggregated approach was suggested for handling tasks of the credit scoring [9-10]. Later it was used for the technical diagnostics of the system state. One and the same classification method is used for plotting of the assembly with compositional approaches (the bagging, the boosting). This method is plotted either at various sample subsets or oriented towards the error compensation of the previous iteration. Multiple use of various classification methods plotted with the learning sample is of interest. In this case to achieve the best result it is necessary to resolve following issues: which learning

methods shall be used? How these methods can be combined? How to make the consistent decision about the operating state of the object based on solutions of certain methods?

We will use the exhaustive enumeration of sets from  $H$  base methods. Then, for example, if  $H = 2$ , we will get three sets: two basic ones and one aggregated; if  $H = 3$ , there will be 7 sets: three basic ones, three aggregated ones, by two basic ones and one aggregated of all three basic methods. It is not too difficult to see that in the general case the number of sets is equal to  $2^H - 1$ . To make the consistent decision about the operating state of the object based on solutions from certain classification methods, we will consider the aggregation of results on the average value, on the median line and using the voting procedure.

Suppose  $\hat{P}_K(X_r)$  is the probability that  $r$ -object is the operating one determined with the aid of the  $K$ -basic method,  $K = 1, \dots, H$ . In mean that when aggregating on the average value:

$$\hat{P}_{AKcp}(X_r) = \frac{\sum_{K=1}^H \hat{P}_K(X_r)}{H}$$

where  $\hat{P}_{AKcp}(X_r)$  is the probability that  $r$ -object is the operating one.

When aggregating on the median line, firstly, it is necessary to range the line with the results of basic methods in the set. If the number of basic methods is odd, the probability that  $r$ -object is the operating one will be:

$$\hat{P}_{AKmed}(X_r) = \frac{\hat{P}_{H+1}(X_r)}{2}$$

If the number of basic methods is even, the relevant probability will be calculated as the half-sum of the median value results.

The result of the aggregated classification method on the voting procedure is the average value of of basic methods results determining the operating state of the object with the probability, for example, not lower than 0.1 ( $\hat{P}_K(X_r) \geq 0,1$ ). Otherwise the probability that  $r$ -object is the operating one is considered as zero.  $\hat{P}_K$  is the probability that  $r$ -object is the operating one at base values of the object  $X_r$  functioning. As can be seen from the above, values of classification probabilities lower than 0.1 are treated as 0 and the rest are treated as 1. Aggregated classification models are plotted using these very values.

In this case, as mentioned above, the division of basic data in the learning sample and the control sample is carried out randomly. That's why structures of aggregated classifiers turn out to be different. The question that has to be answered is what structure to select for making the final decision about the operating state of the object.

As before, we repeated tests 50 times. The sample volume was one and the same (25%) using all eight functioning indicators. Corresponding results of the  $F$ -criterion for five options of every aggregation type are shown in Table 2.

For example, the entry in the first line GrB+DTB+AB means that the aggregate of three basic classifiers including the gradient boosting (GrB), the decision tree bagging (DTB) and the AdaBoost boosting method was the best aggregation option on the average value when using the  $F$ -criterion in this experiment. The number of classifiers included in the aggregate (Table 2) fluctuates from two to six. In the general case it can include all basic classifiers.

Firstly, let us remark that any aggregated method on the  $F$ -criterion turned out to be better than any basic one. Secondly, values of the  $F$ -criterion for aggregated methods are not widely diverging. And finally, it is worth paying attention to the fact that the best of basic methods (the decision tree bagging) is included into the structure of all aggregated classifiers.

It is of interest to study the distribution pattern of the  $F$ -criterion values. As far as the aggregation using the voting procedure is concerned, we applied the following sequence of steps. We used the Statistica system to plot the normal probability curve. Then we transferred this curve to the value distribution histogram (Fig. 2) of this criterion. To check the normality, we used the Shapiro-Wilk criterion recommended for small samples (up to 50 tests). It is apparent that the distribution can be

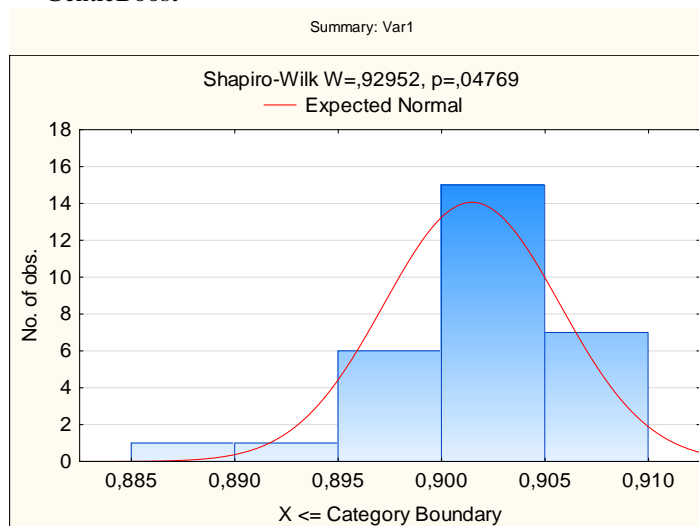
considered as the normal one when the significance level is 0.05. Similar results were obtained for other classifiers (both basic and aggregated) as well.

**Table 2.** F-criterion when aggregating.

Aggregate structure	F-criterion
Aggregation on the average value	
GrB+DTB+AB	0.891
GrB+ DTB	0.889
DTB+AB	0.889
SVM+DTB+AB+LB	0,889
SVM+DTB	0.879
Aggregation on the median line	
DA+SVM+GrB+DTB+AB+GB+RB	0.892
SVM+DTB	0.881
SVM+DTB+AB+RB	0.891
SVM+DTB+LB	0.888
GrB+DTB	0.887
Aggregation using the voting procedure	
NN+SVM+DTB+AB+RB	0.887
GrB+DTB+AB	0.889
DTB+AB	0.889
GrB+DTB	0.885
SVM+GrB+DTB+LB+GB+RB	0.887

Designations:

GrB – the gradient boosting, DTB– the decision tree bagging, AB – AdaBoost, SVM – the support vector machine, LB – LogitBoost, DA – the discriminative analysis, RB – RUSBoost, NN – the neural network, GB – GentleBoost



**Figure 2.** Distribution of F-criterion values.

The distribution normality gives an opportunity to use the standard approach for checking the hypothesis that in the given example the aggregation does lead to the improvement of the diagnostics quality.

We checked the null hypothesis for the equality of F-criterion average values when aggregating and when using basic classification methods (in comparison with data from the decision tree bagging being taken into consideration as the best basic method). As an alternative, we considered the

hypothesis for exceeding of the average value when aggregating. Firstly, we compared dispersions of two samples upon the Fisher criterion (the difference turned out to be statistically non-significant). Then we tried the Student criterion with similar dispersions. It was concluded that the null hypothesis should be rejected: the average value of the  $F$ -criterion when aggregating is higher than when using basic classifiers.

As has already been noted, values of the  $F$ -criterion in Table 2 do not much differ. We checked the hypothesis that the increase in the number of basic classifiers (more than two) in the aggregate structure will non-significantly influence the value of the  $F$ -criterion. We divided the whole sample in two subsets. Data on aggregates consisting of only two components will be included in the first subset. All rest values will be included in the second subset.

Checking of the hypothesis for the equality of average values in these subsets shows its validity: the average value of the  $F$ -criterion does not change when expanding the number of basic classifiers in the aggregate structure.

A consequence of the above result is the fact that it is possible to reduce dramatically the time required for the calculation. Instead of enumerating all aggregation options when searching for the maximum value of the  $F$ -criterion (three aggregation methods and 11 basic classification methods used in the Matlab package,  $3 \cdot (2^{11} - 1) = 6141$  options); it will be enough to enumerate only options including two basic methods ( $3 \cdot 11! / 2! 9! = 165$ ).

It is necessary to take into consideration one more circumstance. During all tests the aggregate included the best basic method (the decision tree bagging). Taking this fact into consideration gives an opportunity to scale back the number of options being enumerated by 30.

However, it is necessary to bear in mind that the given results are obtained in tests of only one technical object. Nevertheless, this experiment shows that the suggested approach shall be approbated for the diagnostics of any other system being studied.

#### 4. Conclusion

To assess the operating state of the object it is recommended to select the simplest aggregated classifier with the sufficiently great value of the  $F$ -criterion. In the given example this classifier is the aggregation on the average value for the decision tree bagging and AdaBoost, or the decision tree bagging together with the gradient boosting (except for sufficiently great values of the  $F$ -criterion, these combinations can be more often found in Table 2).

The considered approach was also used (except for the water treatment system) when assessing the faulty state of the hydroelectric installation on the vibration level and the technological process of the mechanical processing when it showed similar results.

#### 5. References

- [1] Gaskarov D V, Golinkevich T A and Mozgalevskij A V 1974 *Technical condition and reliability prediction of electronic equipment* (Moscow: Soviet radio) p 224
- [2] Klyachkin V N and Buby' D S 2014 Forecasting of technical object state based on piecewise linear regressions *Radioengineering* **7** 137-140
- [3] Krasheninnikov V R, Kuvayskova Yu E, Shunina Yu S and Klyachkin V N 2017 Updating of models predicting objects' state as time series systems and multivariate classifier *Herald of Computer and Information Technologies* **6** 11-16
- [4] Krasheninnikov V R, Klyachkin V N and Kuvayskova Yu. E. 2018 Models updating for technical objects state forecasting *3rd Russian-Pacific Conf. on Computer Technology and Applications (RPC). IEEE Xplore* 1-4.
- [5] Birger I A 1978 *Technical Diagnostics* (Moscow: Engineering) p 240
- [6] Witten I H and Frank E 2005 *Data mining: practical machine learning tools and techniques* (San Francisco: Morgan Kaufmann Publishers) p 525
- [7] Merkov A B 2011 *Pattern recognition. Introduction to statistical learning methods* (Moscow: Editorial URSS) p 256
- [8] Voronina V V, Miheev A V, Yarushkina N G and Svyatov K V 2017 *Machine learning: theory and practice* (Ulyanovsk: UISTU) p 290

- [9] Yumaganov A S and Myasnikov V V 2017 A method of searching for similar code sequences in executable binary files using a featureless approach *Computer Optics* **41(5)** 756-764 DOI: 10.18287/2412-6179-2017-41-5-756-764
- [10] Kropotov Yu A, Proskuryakov A Yu and Belov A A 2018 Method for forecasting changes in time series parameters in digital information management systems *Computer Optics* **42(6)** 1093-1100 DOI: 10.18287/2412-6179-2018-42-6-1093-1100
- [11] Klyachkin V N, Kuvayskova Yu E and Zhukov DA 2017 The use of aggregate classifiers in technical diagnostics, based on machine learning *CEUR Workshop Proc.* **1903** 32-35
- [12] Maksimov A I and Gashnikov M V 2018 Adaptive interpolation of multidimensional signals for differential compression *Computer Optics* **42(4)** 679-687 DOI: 10.18287/2412-6179-2018-42-4-679-687
- [13] Kuvayskova Yu E 2017 The prediction algorithm of the technical state of an object by means of fuzzy logic inference models *Procedia Engineering* **201** 767-772
- [14] Voroncov K V URL: <https://yadi.sk/i/FItIu6V0beBmF>
- [15] Sokolov E A URL: <https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf>
- [16] Davis J and Goadrich M 2006 The relationship between Precision-Recall and ROC curves *Proc. of the 23rd int. conf. on Machine learning* (Pittsburgh) 233-240
- [17] Klyachkin V N and Shunina Yu S 2015 System for borrowers' creditworthiness assessment and repayment of loans forecasting *Herald of Computer and Information Technologies* **11** 45-51
- [18] Neykov M, Jun S Liu and Tianxi Cai 2016 On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting *J. of Machine Learning Research* **17(70)** 1-32
- [19] Wyner A J, Matthew Olson, Justin Bleich and David Mease 2017 Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers *J. of Machine Learning Research* **18(48)** 1-33
- [20] Chen T and Guestrin C 2016 XGBoost: A Scalable Tree Boosting System *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 765-794

### **Acknowledgments**

This test was carried out with the financial support from the Russian Foundation for Basic Research (RFBR) and the Government of Ulyanovsk region, the project 18-48-730001.