Analysis of stable functioning of objects using machine learning

V N Klyachkin[1], D A Zhukov[1] and E A Zentsova[1]

[1]Ulyanovsk State Technical University, Severny Venets street, 32, Ulyanovsk, Russia, 432027

e-mail: v_kl@mail.ru, zh.dimka17@mail.ru

Abstract. Stable functioning of the technical objects is estimated using methods of the statistical process control. However this approach does not always provide the timely detection of violations. It is suggested using machine learning methods for the binary classification of object states (stable or unstable). A program has been developed for calculation in the Matlab environment which allows for analysis of impact of the learning method, classification quality criteria, method of validation set as well as methods of selection of significant indicators on the object's stable functioning forecast precision. Stable operation of the water treatment management system, stable vibration of the hydraulic unit, machining operation process are taken as examples.

1. Introduction

Stable functioning of the technical object is estimated using methods of the statistical process control. However, this approach does not always provide the timely detection of violations [1-2]. It is suggested using machine learning methods for the binary classification of object states (stable or unstable). Alongside with that, the statistical control history can be used as base data for the object state forecasting using machine learning methods. Values of test items and the object operation state are known for every sample.

The binary classification quality depends very heavily on the series of factors. Firstly, it is the selected machine learning method. For example, basic (the naive Bayesian classifier, the neural network and etc.), compositional (various forms of bagging or boosting) and also aggregated methods can be used [3-6]. Secondly, the diagnostic quality depends on the selected criterion. The most commonly used criteria are the error rate in the control sample, *F*-criterion and the area AUC under the ROC-curve [7-8]. Technical objects are characterized by samples with a small volume of non-operating states. In this case the *F*-criterion is the most efficient one. Learning results depend on selection methods of significant criteria of the object operation (using of low-priority criteria can lead to incorrect results), on the forming method of the control sample and on the scope of the control sample [9].

The research objective is to design a program for estimating the stable functioning of the technical object. This program shall analyze the influence of the machine learning method, forming method of the control sample and the selection of significant criteria on the precision of stability forecasting and on carrying out of the numerical analysis of the stable functioning of real objects.

2. Preparation of base data

The statistical process control involves the identification of non-accidental violations connected with so-called special reasons [10-12]. For example, as far as the mechanical processing is concerned these reasons can be the edge wear or its unfastening, composition changes of the cutting fluid and etc.
The main advantage of this approach lies in the fact that violation is identified before the test item oversteps the limits. In addition to the above, statistical methods (control charts are usually used for monitoring of the middle level and the dispersion process) show the violation of its stability (when test features overstep the confidence limit).

Control charts are one of the most efficient tools for the analysis, process monitoring and its preservation in the statistically controlled state. Plotting involves the determination of parameters: volume of the instantaneous sample, time intervals between the sample taking and the position of control limits. Efficiency factors of the control chart are expenditures for carrying out the control procedure; the time interval till the signal collection about the process violation and the probability level of the false alarm.

As far as the multi-parameter process is concerned firstly of all correlation relationships among object performance indicators are examined. Then control facilities are selected and their parameters are specified.

Shewhart charts (when the middle level and its dispersion are controlled described in terms of the range or the standard deviation) are used for independent indicators in accordance with standards. Chart can be combined. For example, charts with average values and chart with ranges; charts with average values and charts with standard deviations, charts with individual observations and charts with moving ranges. The last option is used when measurements are too labor-intensive or expensive. In this case, for the control the result of one measurement is used instead of the instantaneous sample.

Gain in the sensitivity of these charts (reduce the time or the sample volume from the moment when the process violated till the moment when this violation was identified) is possible by using the preventative border and searching for so-called non-accidental structures in the chart. For example, multiple successive increasing or decreasing points indicate the process trend. Multiple points located checker-wise are the indicator of cyclical process fluctuations, etc. Charts of cumulative sums or exponentially weighted moving averages are sometimes used to specify the process stability issue [1-2, 12].

Stability of correlation indicators is estimated using Hotelling algorithms and the generalized variance [13-17]. The Hotelling chart is used to estimate the middle level stability of the multi-variate process. The null-hypothesis of the fact that the mean vector fits the requirements is tested. The generalized dispersion chart is a tool for monitoring of the multi-variate spreading. The hypothesis of the fact that the correlation matrix fits the requirements is tested.

Which controlled indicator or which subrange of indicators is responsible for the process violation? Causes of violations under the multi-variate process control can be identified using the Hotelling partial criterion or the down-weighting of plotted charts. For example, if we control three indicators, then three Hotelling charts shall be plotted for each pair of indicators. This approach is not always correct but in practice it often leads to the identification of the needed indicator. The similar approach can be also used with the generalized variance chart.

Preventative borders , searching for non-accidental structures and charts of multi-variate exponentially weighted moving averages based on both the Hotelling statistics and the generalized variance can be used to improve the efficiency of the multi-variate control.

Let's assume that N samples upon d performance indicators were examined. Control charts identified the stability process violation in k samples. By doing so, the multitude N of precedents is formed $(x^{(i)}, y^{(i)})$, $i = 1 \ldots N$: objects with preset d performance parameters $x = (x_1, x_2, \ldots, x_d)$ and corresponding states y taking one of two possible values (0,1); $y = 0$ corresponds to the unstable state (the number of such precedents is k), $y = 1$ corresponds to the stable state (the number of such precedents is $N - k$). Based on these data it is necessary to restore the dependence between performance indicators and the object state.

3. Object state diagnostics

The implementation of machine learning methods is possible based on the tool library Statistics and Machine Learning Toolbox in the pack Matlab. Taking into consideration research objectives, we designed a program which:

- uses various basic and compositional methods and plots aggregated classifiers of three types: with the aggregation by the average value, by the median or the voting [18-19],
- using various classification quality criteria: error rates in the control sample, *F*-criterion and the area AUC under the ROC-curve,
- selecting significant object performance indicators by plotting the regression model for the dependence of the object state y on performance indicators x_j ($j = 1 \ldots d$) and checking the significance of indicators upon the Student criterion,
- varies the forming method of the control sample (the random selection or the specified base data) and its volume [9].

The base data sample is introduced to diagnose the state stability of the technical object. The machine learning is carried out using all basic and compositional methods integrated in Matlab. In this case, the classification quality is estimated using the cross-validation under the *F*-criterion. Significant performance indicators are selected. Than this procedure is checked whether it improves the quality of the model or not. The sample volume providing the best criterion value is selected by varying the control sample volume from 5 to 25% in comparison with the original sample. Aggregated classifiers are plotted for this very option. The classifier providing the maximum *F*-criterion is selected. If necessary, there is also an opportunity to minimize the error rate in the control sample when plotting aggregated classifiers.

In Fig. 1 shows a flow cyart of the corresponding algorithm.

4. Numerical examination

The statistical control of the mechanical processing stability of the axis (grinding of four cylindrical surfaces) was carried out using 400 samples. All four controlled indicators (axis step diameters) turned out to be correlated with each other. The Hotelling chart was used to control the middle level of the process. The generalized variance chart was used to control the multi-variate dispersion. Fig. 2 shows the results of the statistical control for the first 51 samples. The critical value of the Hotelling statistics was 10.830. The critical value of the generalized variance was $0.45*10^{-8}$. Charts were plotted using the system Statistica.

It is apparent that process violations (overstepping the limits) in the Hotelling chart happened in three samples: 38, 45 and 47. Process violations in the generalized variance chart happened in two samples: 33 and 34. As can be seen from the above, violations of the controlled process happened in 5 cases of 51 precedents.

Table with base data was prepared upon all 400 observations. When carrying out the examination using the designed program it was found that all four indicators were significant. The maximum value of the *F*-criterion (0.874) was measured for the control sample of the volume 15% in comparison with the original sample when using the aggregated classifier by the average value including the support vector machine and the decision tree bagging. In addition to the above, the value of the *F*-criterion was boosted by 8%. The best classifier by separate methods (the decision tree bagging) was considered as the standard approach.

During the second test the vibration stability of the hydraulic aggregate was estimated by values of 10 gauges in 5000 observations [20]. The multitude of 10 values was divided into four sub-multitudes. Two values turned out to be uncorrelated with others. Three correlated values formed the third group. Five correlated values formed the forth group. First two independent values were controlled using Shewhart charts for average values and standard deviations. The Hotelling and the generalized variance charts were plotted for the third and the forth groups. According to the results of the statistical control the sample of base data was formed. The machine learning was carried out using this sample. Seven indicators out of ten turned out to be significant. The aggregated classifier by the median including the gradient boosting and the logistic regression turned out to be the best. The

volume of the control sample was 20%. The increase of the *F*-criterion (up to the value 0.904) was 18%.

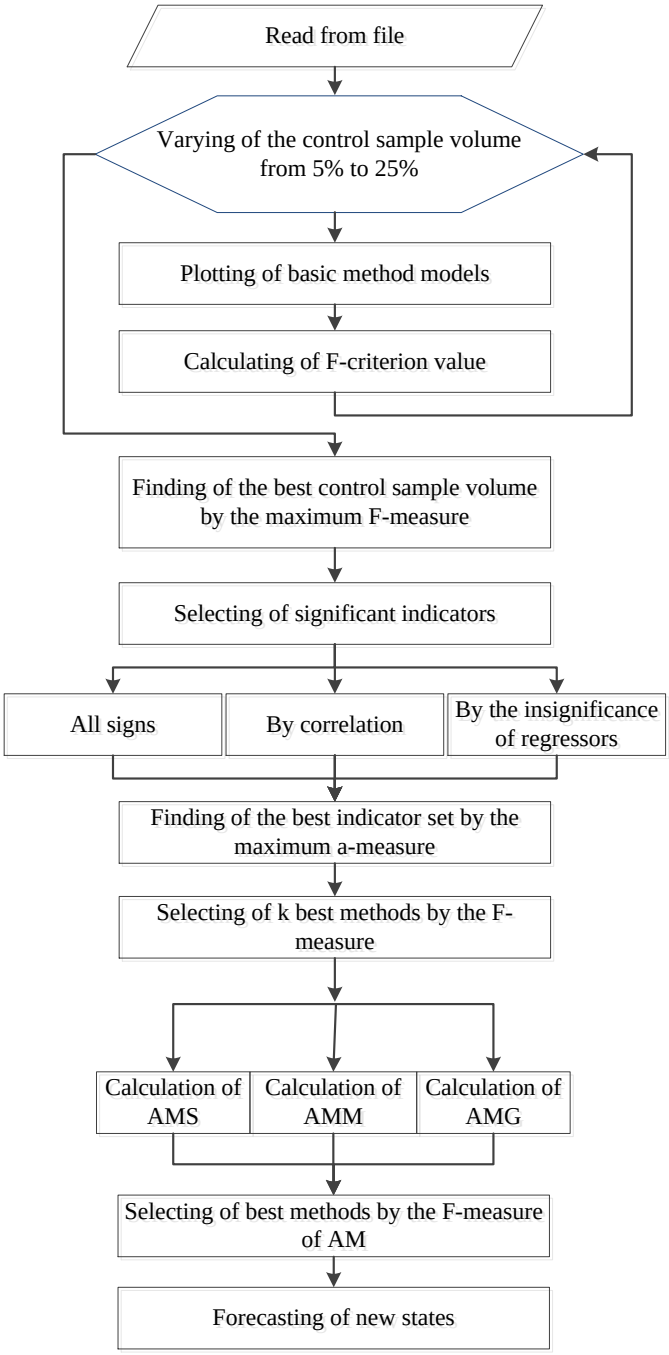


Figure 1. Flow chart of the algorithm for stability forecasting.

When analyzing the stable functioning of the water purification system by eight quality indicators of the potable water results of 1557 observations (operating state was registered in 1204 cases) were used. Six indicators turned out to be significant. Maximum value of the *F*-criterion was when aggregating the neural network and the decision tree bagging. During this test the increase of the criterion value in comparison with the neural network (the best separate classifier) was negligible: from 0.879 to 0.881. In addition to the above, the volume of the control sample was 10%.

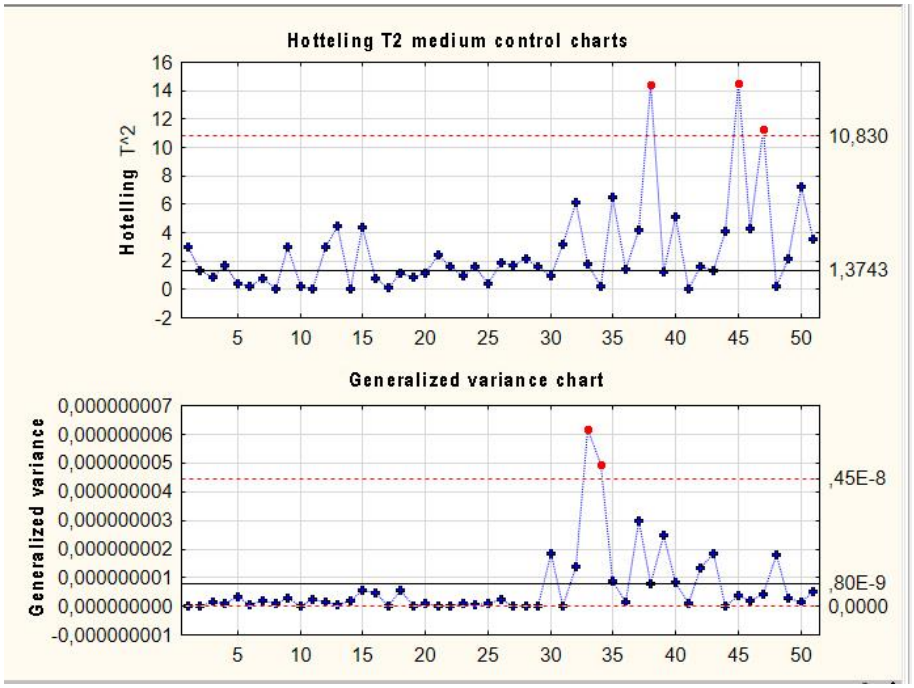
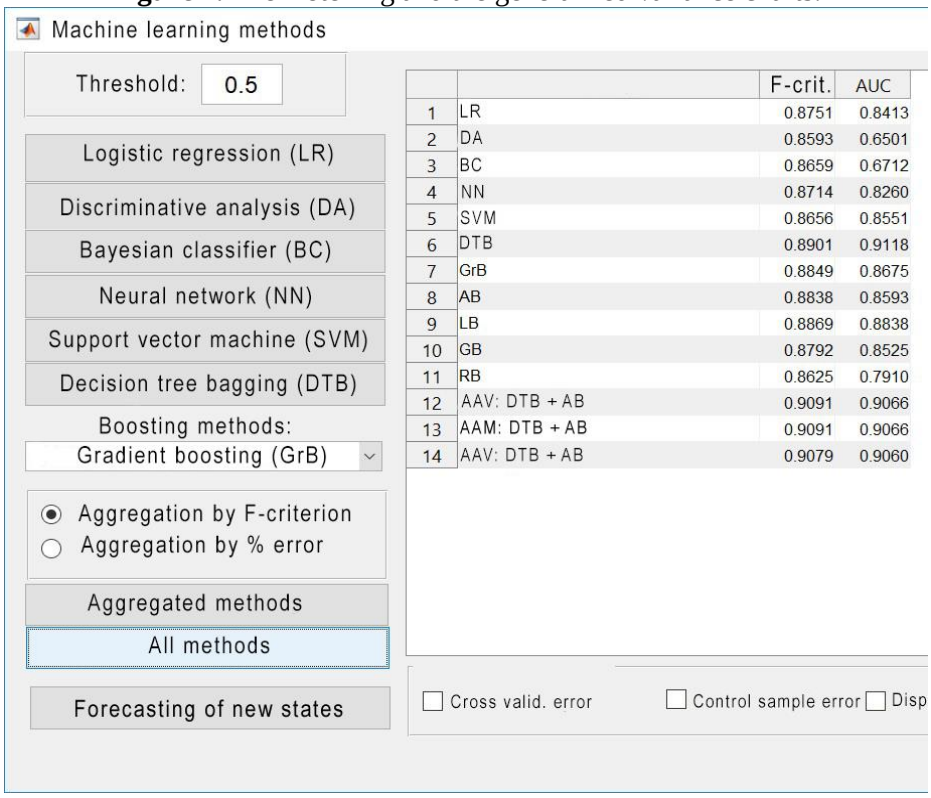


Figure 2. The Hotelling and the generalized variance charts.



		F-crit.	AUC
1	LR	0.8751	0.8413
2	DA	0.8593	0.6501
3	BC	0.8659	0.6712
4	NN	0.8714	0.8260
5	SVM	0.8656	0.8551
6	DTB	0.8901	0.9118
7	GrB	0.8849	0.8675
8	AB	0.8838	0.8593
9	LB	0.8869	0.8838
10	GB	0.8792	0.8525
11	RB	0.8625	0.7910
12	AAV: DTB + AB	0.9091	0.9066
13	AAM: DTB + AB	0.9091	0.9066
14	AAV: DTB + AB	0.9079	0.9060

Figure 3. Calculation results.

Fig. 3 shows the program window with one calculation options. It is apparent that in this very case the value of the *F*-criterion for basic binary classification methods falls in the range from 0.8593 for the discriminative analysis and up to 0.8901 for the decision tree bagging. All aggregated methods have one and the same structure (this happens by no means always) and include two basic methods: the decision tree bagging and AdaBoost. In the general case the number of basic methods can be even greater. Please note that the best of basic classifiers (the decision tree bagging) is almost always

included in the structure of the aggregated classifier. The value of the *F*-criterion by aggregated methods is some greater then by any basic classifiers. For the aggregation by the average value and by the median it is equal and amounts 0.9091.

Alongside with that, the decision tree bagging showed the greatest AUC - area under the error curve (0.9118). Like for the *F*-criterion, the closer is this value to 1, the higher is the classification quality. The maximum value by aggregated methods (0.9066) turned out to be somehow less.

5. Conclusion

The carried out research showed that the stable functioning analysis of the technical object can be carried out both using methods of the statistical process control, widely used during the technological process control, and using the binary classification of object states based on the machine learning. Previous results of the statistical control can be used as base data and precedents.

Shewhart charts are used for the middle level process control and its dispersion when exercising control over independent indicators of the object operation. The stability of correlated indicators is estimated using the Hotelling and the generalized variance algorithms. The Hotelling chart is used to estimate the middle level stability of the multi-variate process. The generalized variance chart is used as the tool for monitoring of the multi-variate dispersion.

The base data sample is introduced to diagnose the state stability of the technical object using the machine learning. Learning of the binary classification model is carried out using all basic and compositional methods integrated in Matlab. In this case, the classification quality is estimated using the cross-validation under the *F*-criterion. The diagnostics precision can be increased due to the aggregation of methods, the selection of significant indicators and changing of the control sample volume.

The designed program provides the automatic searching for the best diagnostic option of the object state by the preset criterion. As examples, the stable functioning of the water purification control system, the hydraulic aggregate vibration stability and the technological procedure of mechanical processing were considered.

6. References

[1] Klyachkin VN and Karpunina I N 2018 Statistical methods for assessing the stability of functioning of technical systems *Reliability and Quality of Complex Systems* **2** 36-42
[2] Klyachkin V N 2008 The static control system for multivariate manufacturing process *Instruments and Systems: Monitoring, Control, and Diagnostics* **10** 30-33
[3] Witten I H and Frank E 2005 *Data mining: practical machine learning tools and techniques* (San Francisco: Morgan Kaufmann Publishers) p 525
[4] Merkov A B 2011 *Pattern recognition.Introduction to statistical learning methods* (Moscow: Editorial URSS) p 256
[5] Voronina V V, Miheev A V, Yarushkina N G and Svyatov K V 2017 *Machine learning: theory and practice* (Ulyanovsk: UlSTU) p 290
[6] Voroncov K V URL: https://yadi.sk/i/FItIu6V0beBmF
[7] Sokolov E A URL: https://github.com /esokolov/ml-course-hse/blob/master/2018-fall/lecture-notes/lecture04-linclass.pdf
[8] D'yakonov A M URL: https:// dyakonov.org/2017/07/28/auc-roc-ploshchad'-pod-krivoj-oshibok/#more-5362
[9] Zhukov D A and Klyachkin V N 2018 The Effect of the Control Sample Volume on the Quality of Diagnostics of the Technical Object State *Automation of Control Processes* **2** 90-95
[10] Wheeler D and Chambers D 1992 *Understanding Statistical Process Control* (SPC Press) p 409
[11] Montgomery D C 2009 *Introduction to statistical quality control* (New York: John Wiley and Sons) p 754
[12] Maksimov A I and Gashnikov M V 2018 Adaptive interpolation of multidimensional signals in differential compression *Computer Optics* **42(4)** 679-687
[13] Yumaganov A S and Myasnikov V V 2017 The method of searching for similar code sequences in executable binary files using the unmarked approach *Computer Optics* **41(5)** 756-784

[14] Klyachkin V N and Bubyr' D S 2014 Forecasting of technical object state based on piecewise linear regressions *Radioengineering* **7** 137-140
[15] Klyachkin V N and Karpunina I N 2017 The analysis of technical object functioning stability as per the criterion of monitored parameters multivariate dispersion *CEUR Workshop Proc.* **1903** 28-31
[16] Salmasnia A, Kaveie M and Namdar M 2018 An integrated production and maintenance planning model under VP-T2 Hotelling chart *Computers & Industrial Engineering* **18** 89-103
[17] Franceschini F, Galetto M and Genta G 2015 Multivariate control charts for monitoring internal camera parameters in digital photogrammetry for LSDM (Large-Scale Dimensional Metrology) applications *Precision Engineering* **42** 133-142
[18] Klyachkin V N, Kuvayskova Yu E and Zhukov D A 2017 The use of aggregate classifiers in technical diagnostics, based on machine learning *CEUR Workshop Proc.***1903** 32-35
[19] 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
[20] Kropotov Yu A, Proskuryakov A Yu and Belov A A 2018 A method for predicting changes in the parameters of time series in digital information control systems *Computer Optics* **42(6)** 1083-1100

Acknowledgments
This surveying 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 should be used.