**Серия «Математика»**
2021. Т. 38. С. 84—95

Онлайн-доступ к журналу:
http://mathizv.isu.ru

ИЗВЕСТИЯ
Иркутского
государственного
университета

УДК 519.246
MSC 68T10, 62H30
DOI https://doi.org/10.26516/1997-7670.2021.38.84

# On the Accuracy of Cross-Validation in the Classification Problem*

V. M. Nedel'ko
*Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russian Federation*

**Abstract.** In this work we will study the accuracy of the cross-validation estimates for decision functions. The main idea of the research consists in the scheme of statistical modeling that allows using real data to obtain statistical estimates, which are usually obtained only by using model (synthetic) distributions.

The studies confirm the well-known empirical recommendation to choose the number of folds equal to 5 or more. The choice of more than 10 folds does not yield a significant increase in accuracy. The use of repeated cross-validation also does not provide fundamental gain in precision.

The results of the experiments allow us to formulate an empirical fact that the accuracy of the estimates obtained by the cross-validation method is approximately the same as the accuracy of the estimates obtained from the test sample of half the size. This result can be easily explained by the fact that all the objects of the test sample are independent, and the estimates built by the cross-validation on different subsamples (folds) are not independent.

**Keywords:** K-fold cross-validation, accuracy, statistical estimates, machinelearning.

## 1. Introduction

The K-fold cross-validation method is the most commonly used method of evaluating the quality of solutions [6;7] in machine learning problems [5; 11; 12; 17]. However, despite the large number of papers devoted to the

* The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no 0314-2019-0015) and with partial support by RFBF grant 19–29–01175.

study of this approach, the problem of assessing the accuracy of the resulting quality estimates remains open [10;13;14]. In particular, confidence intervals for the cross-validation are currently unknown, there are only very rough estimates of such intervals [19].

The paper [18] shows that the formally constructed 95% confidence interval for cross-validation estimates is not even an estimate of the confidence interval, since the probability of going beyond the interval is much greater than 0,05. At the same time, in the mentioned work the term confidence interval, however, continues to be used, as in many other sources. But in the strict sense, such constructions shouldn't be called as confidence intervals.

In [1] the confidence intervals based on CV-score are proposed, but those intervals estimate some non-standard version of test error.

The question of choosing the optimal number of folds $K$ is also actual [2]. In some sources there is a recommendation to choose $K = 10$ and carry out 10 repetitions of the cross-validation procedure. Other authors believe that it is advisable to choose $K$ as large as the computational capabilities (and sample size) allow.

The goal of this work is to develop some empirical approach for investigation of the accuracy of cross-validation. This approach includes the method of statistical modelling on real data that allows to introduce and estimate some new statistical characteristics. In particular, this allows to substantiate the choice of number of folds.

## 2. Problem statement

Consider the statistical formulation of the problem of decision function construction [3; 15; 16].

Let $X$ be the set (space) of values of a given set of variables $x_1, ..., x_n$, and $Y$ is the set of values of the target variable. Suppose that a probability space ($\sigma$-algebra and probability measure $\mathsf{P}$) is given on $X \times Y$.

We introduce a notation for the expectation of a measurable function $f : X \times Y \to R$

$$\mathsf{E}_{X,Y} f(x, y) = \int_{X \times Y} f(x, y)\ \mathsf{P}(dx, dy).$$

The decision function is a mapping $\lambda : X \to \mathfrak{S}$, where $\mathfrak{S}$ is the set of solutions (the space of estimates). The values of $\mathfrak{S}$ can be, in particular, the values of the variable $Y$ or, for example, estimates of the probability that an object belongs to a given class.

The quality of the decision function is estimated by some criterion $\mathcal{K}(\lambda, \mathsf{P})$, which in general is a functional of the decision function and distribution.

In this paper we will consider quality criteria that can be represented as a mathematical expectation from the so-called loss function $L : \mathfrak{S} \times Y \to R$, i.e.

$$\mathcal{K}(\lambda, \mathsf{P}) = \mathsf{E}_{X,Y} L(\lambda(x), y)$$

Some criteria (for example, the area under the error curve) can not be represented in this form, but most of the quality criteria used in practice can be expressed through the loss function.

The problem of constructing the decision function is to construct $\lambda$, which would minimize the criterion $\mathcal{K}(\lambda, \mathsf{P})$. Particular statements of this problem are the problems of classification and regression analysis (recovering of dependencies) [4] [8] [9].

In this formulation, the problem is actually not yet set, since the criterion depends on the distribution, which in real problems is unknown.

In practice, a decision function is constructed based on a sample

$$V_N = \left((x_i, y_i), i = 1, ..., N\right), \;\; x_i \in X, \; y_i \in Y, \;\; V_N \in W_N, \; W_N = (X \times Y)^N.$$

Method of constructing the decision functions is a mapping

$$Q : W \to \Lambda, \quad W = \bigcup_{N=1}^{\infty} W_N,$$

where $\Lambda$ is a given class of decision functions.

The purpose of this work is to evaluate the quality of the solution constructed by a given method.

Let $\lambda_{Q,V_N} \equiv Q(V_N)$ be a solution built by the method $Q$ on the sample $V_N$.

We are interested in the value

$$\mathcal{K}_Q(V_N) = \mathsf{E}_{X,Y} L(\lambda_{Q,V_N}(x), y).$$

Here we have introduced the notation $\mathcal{K}_Q(V_N)$ instead of $\mathcal{K}(\lambda_{Q,V_N}, \mathsf{P})$ to simplify the writing, and to emphasize that the criterion is a function of sample, while the parameters $Q$ and $\mathsf{P}$ will be fixed.

Because the expression is a function of sample, it is a random variable.

In real machine learning tasks we have some fixed sample. If the sample is fixed, than $K_Q(V_N)$ becomes the unknown value dependent only on the distribution, which can be evaluated by statistical methods similar to the estimation of parameters of distributions.

For such evaluation, the method of cross-validation is usually used.

In this method, the sample is split into $K$ disjoint parts of $V_N^1, ..., V_N^K$ of equal (or nearly equal) size.

A cross-validation estimate is the value

$$\widetilde{\mathcal{K}}_{Q,K}(V_N) = \frac{1}{N} \sum_{i=1}^{N} L(\lambda_{Q,V_N \setminus V_N^{k(i)}}(x_i), y_i).$$

Here $V_N^{k(i)}$ is the part of the sample that contains $(x_i, y_i)$. This fold is removed from the original sample when you build a solution for $(x_i, y_i)$.

The error of a cross-validation estimate is expressed by the value

$$\Delta(V_N) = \widetilde{\mathcal{K}}_{Q,K}(V_N) - \mathcal{K}_Q(V_N).$$

The main characteristic of cross-validation accuracy will be the standard error

$$err_{CV} = \sqrt{\mathsf{E}_{W_N}(\Delta(V_N))^2}.$$

The expectations are taken over all samples of size $N$.

Let us use the decomposition

$$\mathsf{E}_{W_N}(\Delta(V_N))^2 = \tilde{\sigma}^2 + \sigma^2 + bias_{CV}^2 - 2\tilde{\sigma}\sigma\varkappa,$$

where

$$\tilde{\sigma}^2 = \mathsf{D}_{W_N}\widetilde{\mathcal{K}}_{Q,K}(V_N), \qquad \sigma^2 = \mathsf{D}_{W_N}\mathcal{K}_Q(V_N),$$

$$bias_{CV} = \mathsf{E}_{W_N}\Delta(V_N), \qquad \varkappa = corr_{W_N}(\widetilde{\mathcal{K}}_{Q,K}(V_N), \mathcal{K}_Q(V_N)).$$

Here $corr_{W_N}$ is a normalized correlation coefficient (the expectation is taken over $W_N$), $\mathsf{D}$ denotes a variance.

The goal of this work is to investigate properties of $err_{CV}$ and to propose a method to estimate it.

## 3. Proposed statistics

Now let us introduce a notation for the standard deviation of the loss

$$\mathcal{S}_Q(V_N) = \sqrt{\mathsf{D}_{X,Y}L(\lambda_{Q,V_N}(x), y)}.$$

The distribution on the value of losses depends on the constructed decision function, and hence on the training sample. To characterize how significant this dependence is, we introduce the value

$$\varepsilon_S = \sqrt{\mathsf{D}_{W_N} \ln \mathcal{S}_Q(V_N)}.$$

The logarithm is used to get a scale-free measure that is a characteristic of the relative spread of a random variable (given its positivity).

The value $\mathcal{S}_Q(V_N)$ can not be calculated directly, so we will estimate it via

$$(\widetilde{\mathcal{S}}_Q(V_N))^2 = \frac{1}{N}\sum_{i=1}^{N}\left(L\left(\lambda_{Q,V_N\setminus V_N^{k(i)}}(x_i), y_i\right) - \widetilde{\mathcal{K}}_{Q,K}(V_N)\right)^2.$$

The relative deviation of the obtained estimate is defined as

$$\tilde{\varepsilon}_S = \sqrt{\mathsf{D}_{W_N} \ln \widetilde{\mathcal{S}}_Q(V_N)}.$$

If we have a test sample, we can estimate $\mathcal{K}_Q(V_N)$ with inaccuracy of order $\frac{\mathcal{S}_Q(V_N)}{\sqrt{N}}$.

Experimental studies show that the error of cross-validation estimates is much higher than the estimate obtained from the test sample of the same size.

In our experiments we will investigate a possibility to estimate $err_{CV}$ on the base of $\widetilde{\mathcal{S}}_Q(V_N)$.

## 4. Method of research

The idea of the approach is based on the fact that the estimates made for a large sample are close to the expected values (i.e. to the values obtained on the distributions themselves). Note that in our case there is some subtlety, namely that the notion of a large sample becomes relative. To study the properties of quality estimates, it is important that the size of the test sample will be much larger than the size of training samples.

The scheme of the experimental study is as follows. The initial sample is divided into two approximately equal parts. One part is reserved for the test. The second part is divided into many training samples, each of which is used for independent training, as well as for evaluation by cross-validation.

This approach allows us to draw conclusions with same statistical reliability as by modeling on distributions, but based on real data.

## 5. Experimental results

Numerical studies was performed on the task "adults" from the UCI repository. The data size is 32560 object with 5 variables (we selected only numeric features). The target variable is represented by the values of two classes.

We used logarithmic loss function

$$L(\lambda(x), y) = -y \ln \lambda(x) - (1 - y) \ln(1 - \lambda(x)),$$

where $y \in \{0, 1\}$, $\lambda(x) \in (0, 1)$.

To avoid too large losses, the predicted probabilities were clipped to a range $[0,001, 0,999]$.

We used gradient boosting as a classification method (class GradientBoostingClassifier from sklearn.ensemble), max_depth=2.

Table 1

The parameters of the experiment

| N | n_estimators | learning_rate | $\mathsf{E}\overline{\mathcal{K}}_Q$ | $\mathsf{E}\mathcal{K}_Q$ | $\sigma$ | $\mathsf{E}\mathcal{S}_Q$ | $\varepsilon_S$ | $err_{test}$ | Tab. |
|---|---|---|---|---|---|---|---|---|---|
| 1000 | 100 | 0,2 | 0,314 | 0,421 | 0,007 | 0,637 | 0,064 | 0,019 | 2, 3 |
| 100 | 30 | 0,05 | 0,368 | 0,480 | 0,020 | 0,553 | 0,175 | 0,056 | 4, 5 |

Table 2

Characteristics of the cross-validation estimate when $N = 1000$

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{\mathcal{S}}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,040 | 0,052 | 0,030 | -0,421 | 0,729 | 0,051 | 0,025 | 1,367 |
| 3 | 0,022 | 0,035 | 0,024 | -0,390 | 0,690 | 0,045 | 0,028 | 0,642 |
| 5 | 0,012 | 0,032 | 0,026 | -0,454 | 0,662 | 0,039 | 0,042 | 0,397 |
| 10 | 0,007 | 0,031 | 0,027 | -0,485 | 0,649 | 0,036 | 0,065 | 0,250 |

The table 1 contains parameters (training sample size $N$, n_estimators, learning_rate) , the experimental results (those doesn't depend on the number of folds $K$), and the references on the tables with corresponding results.

The column $\mathsf{E}\overline{\mathcal{K}}_Q(V_N)$ contains the average values of the quality criterion on the training sample.

The values of $\sigma$ are relatively small, which means that the quality of solutions does not depend very much on a particular training sample.

The $err_{test}$ column contains quality criterion estimates based on a test sample of the same size as the training sample.

The table 2 presents the results of experiments to evaluate the characteristics of the cross-validation depending on $K$.

You can see that the bias of the cross-validation estimate becomes insignificant at $K = 10$, i.e. too small in comparison with the variance.

The error of cross-validation estimation at small $K$ decreases slightly with the growth of $K$, at $K > 10$ the error varies slightly.

The accuracy of the cross-validation estimation is significantly lower than the accuracy of the $err_{test}$ estimate by the test sample of the same size.

The value $S_{folds}$ is the standard deviation of the cross-validation score over $K$ folds. Some sources recommend to build an estimate of the accuracy of the cross-validation based on $S_{folds}$. However, the large values of $R_{folds}$, which are the standard deviation of $S_{folds}$, indicate that from sample to sample the values of $S_{folds}$ can vary significantly, so any estimates based on them will be unreliable.

Table 3

Characteristics of the cross-validation when $N = 1000$, number of repeats is 10

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{\mathcal{S}}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,034 | 0,046 | 0,029 | -0,430 | 0,730 | 0,031 | 0,019 | 1,328 |
| 3 | 0,019 | 0,037 | 0,029 | -0,563 | 0,686 | 0,029 | 0,025 | 0,699 |
| 5 | 0,010 | 0,034 | 0,029 | -0,505 | 0,662 | 0,034 | 0,037 | 0,326 |
| 10 | 0,005 | 0,031 | 0,027 | -0,446 | 0,648 | 0,033 | 0,054 | 0,272 |

Table 4

Characteristics of the cross-validation estimate when $N = 100$

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{\mathcal{S}}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,030 | 0,092 | 0,082 | -0,137 | 0,628 | 0,159 | 0,054 | 1,266 |
| 3 | 0,013 | 0,081 | 0,076 | -0,085 | 0,590 | 0,127 | 0,077 | 0,771 |
| 5 | 0,000 | 0,072 | 0,069 | -0,045 | 0,563 | 0,111 | 0,107 | 0,385 |
| 10 | -0,006 | 0,071 | 0,068 | -0,028 | 0,550 | 0,106 | 0,166 | 0,255 |
| 20 | -0,007 | 0,071 | 0,068 | -0,004 | 0,546 | 0,104 | 0,240 | 0,190 |

The table 3 presents the results of similar experiments in which the procedure of cross-validation is repeated 10 times for random permutations of sample objects.

As you can see, the accuracy of the estimates increases slightly, and only for small $K$.

Tables 4 and 5 present the results of similar experiments when $N = 100$.

As expected, the accuracy of estimates decreases, but the qualitative conclusions are the same.

For comparison, similar calculations were performed for the BNP Paribas Cardif Claims Management problem (https://www.kaggle.com/c/bnp-paribas-cardif-claims-management).

We kept only numerical variables with the least number of missing values: v10, v12, v14, v21, v34, v40, v50, v114. The parameters of the algorithm: n_estimators: 100, learning_rate: 0,1, max_depth: 2.

The results are given in tables 6, 7. There is no qualitative difference from previous experiments.

Since $\tilde{\varepsilon}_S$ is small, the $\widetilde{\mathcal{S}}_Q(V_N)$ may be used as an estimate for $\mathcal{S}_Q(V_N)$.

Finally, we can estimate $err_{CV}$ by the value $\frac{\widetilde{\mathcal{S}}_Q(V_N)}{\sqrt{N/2}}$. This is rough and fully heurisric estimate, but it seems that we have nothing better.

Table 5

Characteristics of the cross-validation when $N = 100$, number of repeats is 10

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{S}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,024 | 0,085 | 0,077 | -0,117 | 0,628 | 0,102 | 0,048 | 0,989 |
| 3 | 0,016 | 0,076 | 0,071 | -0,049 | 0,587 | 0,097 | 0,080 | 0,691 |
| 5 | -0,001 | 0,070 | 0,067 | -0,029 | 0,565 | 0,100 | 0,103 | 0,464 |
| 10 | -0,004 | 0,069 | 0,066 | -0,017 | 0,553 | 0,103 | 0,165 | 0,317 |

Table 6

Characteristics of the cross-validation when $N = 1000$, dataset "Paribas"

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{S}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,020 | 0,032 | 0,024 | -0,234 | 0,633 | 0,039 | 0,019 | 1,055 |
| 3 | 0,009 | 0,022 | 0,020 | -0,059 | 0,605 | 0,037 | 0,028 | 0,522 |
| 5 | 0,005 | 0,021 | 0,020 | -0,233 | 0,591 | 0,033 | 0,038 | 0,378 |
| 10 | 0,001 | 0,020 | 0,019 | -0,210 | 0,582 | 0,035 | 0,056 | 0,286 |

## 6. Conclusion

The studies confirm the well-known empirical recommendation to choose the number of partitions (folds) in cross-validation equal to 5 or 10. At $K = 5$, the accuracy of the estimates is slightly lower, but in practice it is usually not significant, so $K = 5$ is also a justified choice.

Of greater interest are the obtained results on the accuracy of the estimates. The results of the experiments allow us to formulate an empirical assessment that the accuracy of the estimates obtained by cross-validation is approximately the same as the accuracy of the estimates obtained from the test sample of half the size.

The fact that cross-validation estimate is much less accurate that hold-out one can be qualitatively explained by that all the objects of the test sample are independent, and the estimates built by the cross-validation on different folds are not independent.

For evaluating the accuracy of cross-validation estimates one should evaluate the variance of the values of the loss function calculated separately for each object of the sample (using the decision function constructed without using this object). The resulting value should be divided by $N/2$. This will give a rough estimate of the variance of the quality estimated by the cross-validation.

Table 7

Characteristics of the cross-validation when $N = 1000$, number of repeats is 10

| K | $bias_{CV}$ | $err_{CV}$ | $\tilde{\sigma}$ | $\varkappa$ | $\mathsf{E}_{W_N}\widetilde{S}_Q(V_N)$ | $\tilde{\varepsilon}_S$ | $S_{folds}$ | $R_{folds}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0,027 | 0,032 | 0,018 | -0,283 | 0,638 | 0,028 | 0,024 | 0,191 |
| 3 | 0,016 | 0,022 | 0,015 | -0,266 | 0,607 | 0,033 | 0,030 | 0,142 |
| 5 | 0,010 | 0,020 | 0,017 | -0,320 | 0,591 | 0,034 | 0,038 | 0,127 |
| 10 | 0,007 | 0,018 | 0,016 | -0,319 | 0,582 | 0,036 | 0,054 | 0,100 |

## References


1. Bayle P., Bayle A., Janson L., Mackey L. Cross-validation Confidence Intervals for Test Error. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 16339-16350.
2. Beleites C., Baumgartner R., Bowman C., Somorjai R., Steiner G., Salzer R., Sowa M. G. Variance reduction in estimating classication error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems*, 2005, vol. 79, iss. 1-2, pp. 91-100, https://doi.org/10.1016/j.chemolab.2005.04.008
3. Franc V., Zien A., Schölkopf B. Support Vector Machines as Probabilistic Models. *Proc. of the International Conference on Machine Learning (ICML).* ACM, New York, USA, 2011, pp. 665-672.
4. Friedman J., Hastie T., Tibshirani R. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 2000, vol. 28, pp. 337-407. https://doi.org/10.1214/aos/1016218223
5. Kelmanov A.V., Pyatkin A.V. NP-trudnost nekotorykh kvadratichnykh evklidovykh zadach 2-klasterizatsii [NP-hardness of some quadratic Euqleadean biclasterization tasks]. *Doklady Akademii Nauk* [Reports of Academy of Science], 2015, vol. 464, no. 5, pp. 535-538. https://doi.org/10.7868/S0044466916030091 (in Russian)
6. Lbov G. S., Starceva N. G. Sravnenie algoritmov raspoznavanija s pomoshh'ju programmnoj sistemy "Poligon" [Comparison of recognition algorithms with the software system "Poligon"]. *Analiz dannyh i znanij v jekspertnyh sistemah* [Analysis of data and knowledge in expert systems], Novosibirsk, 1990, iss. 134, Vychislitel'nye sistemy [Computer systems], pp. 56-66. (in Russian)
7. Lbov G. S., Starceva N. G. *Logicheskie reshajushhie funkcii i voprosy statisticheskoj ustojchivosti reshenij* [Logical decision functions and problem of statistical robustness of the solutions]. Novosibirsk, Institute of Mathematics SB RAS Publ., 1999, 211 p. (in Russian)
8. Lugosi G., Vayatis N. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 2004, vol. 32, pp. 30-55. https://doi.org/10.1214/aos/1079120129
9. Mease D., Wyner A. Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 131-156. https://doi.org/10.1145/1390681.1390687
10. Motrenko A., Strijov V., Weber G.-W. Sample Size Determination For Logistic Regression. *Journal of Computational and Applied Mathematics*, 2014, vol. 255, pp. 743-752. https://doi.org/10.1016/j.cam.2013.06.031

11. Krasotkina O.V., Turkov P.A., Mottl V.V. Bayesian Approach to the Pattern Recognition Problem in Nonstationary Environment. *Lecture Notes in Computer Science*, 2011, vol. 6744, pp. 24-29. https://doi.org/10.1007/978-3-642-21786-9_6.
12. Krasotkina O.V., Turkov P.A., Mottl' V.V. Bajesovskaja logisticheskaja regressija v zadache obuchenija raspoznavaniju obrazov pri smeshhenii reshajushhego pravila [Bayesian logistic regression in the problem of pattern recognition learning on shifting decision rule]. *Izvestija Tulskogo gosudarstvennogo universiteta. Tehnicheskie nauki.* [Proceedings of the Tula State University. Engineering,] 2013, no. 2, pp. 177-187. (in Russian)
13. Nedel'ko V.M. Misclassification probability estimations for linear decision functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2004, vol. 3138, pp. 780-787. https://doi.org/10.1007/978-3-540-27868-9_85
14. Nedel'ko V. Decision trees capacity and probability of misclassification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. LNAI, 2005, vol. 3505, pp. 193-199. https://doi.org/10.1007/11492870_16
15. Nedel'ko V. M. Regressionnye modeli v zadache klassifikacii [Regression models in the classification problem]. *Sibirskij zhurnal industrialnoj matematiki* [Siberian Journal of Industrial Mathematics], 2014, vol. 27, no. 1, pp. 86-98. (in Russian)
16. Nedel'ko V. M. K voprosu ob jeffektivnosti bustinga v zadache klassifikacii [On the boosting efficiency in the classification problem]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Serija: matematika, mehanika, informatika*. [Bulletin of the Novosibirsk State University. Series: Mathematics, Mechanics, Computer Science], 2015, vol. 15, iss. 2, pp. 72—89. (in Russian), https://doi.org/10.17377/PAM.2015.15.206
17. Torshin I.Yu., Rudakov K.V. On the Theoretical Basis of Metric Analysis of Poorly Formalized Problems of Recognition and Classification. *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications),* 2015, vol. 25, no. 4, pp. 577-587. https://doi.org/10.1134/S1054661815040252
18. Vanwinckelen G., Blockeel H. On estimating model accuracy with repeated cross-validation. *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, 2012, pp. 39-44.
19. Vorontsov K.V. Exact Combinatorial Bounds on the Probability of Overfitting for Empirical Risk Minimization. *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)*, 2010, vol. 20, no. 3, pp. 269-285, https://doi.org/10.1134/S105466181003003X

**Victor Nedel'ko**, Candidate of Sciences (Physics and Mathematics), Sobolev Institute of Mathematics SB RAS, 4, Koptjuga, Novosibirsk, 630090, Russian Federation, tel.: +7(383)3332793,
email: nedelko@math.nsc.ru,
ORCID iD https://orcid.org/0000-0001-8217-9304

*Received 30.10.2021*

---

## О точности оценок скользящего экзамена в задаче классификации

В. М. Неделько

*Институт математики им. С. Л. Соболева СО РАН, Новосибирск, Российская Федерация*

**Аннотация**. Метод скользящего экзамена (K-fold cross-validation) является наиболее часто используемым методом оценивания качества решений в задачах машинного обучения. Несмотря на большое число работ, посвященных исследованию данного подхода, остается открытой проблема оценивания точности получаемых оценок качества. В частности, в настоящее время неизвестны доверительные интервалы для оценки скользящего экзамена, существуют лишь очень грубые оценки таких интервалов.

Основной идеей работы является схема статистического моделирования, которая позволяет использовать реальные данные для получения статистических оценок, которые обычно получаются только при использовании модельных распределений. Предложенный подход позволяет достаточно точно вычислять как общую погрешность оценок скользящего экзамена, так и отдельные ее компоненты (смещение, дисперсию), а также оценивать связь этой погрешности с различными статистиками.

Использование повторяющегося скользящего экзамена со случайным разбиением на подвыборки также не дает принципиального выигрыша в точности. Результаты экспериментов позволяют сформулировать эмпирическую оценку, что точность оценок, полученных методом скользящего экзамена приблизительно такая же, как точность оценок, полученных по контрольной выборке, вдвое меньшего объема. Этот результат легко объяснить тем фактом, что все объекты контрольной выборки независимы, а оценки, построенные скользящим экзаменом на разных подвыборках, не являются независимыми.

**Ключевые слова:** построение решающих функций, скользящий экзамен, точность статистических оценок, машинное обучение.

## Список литературы

1. Bayle P., Bayle A., Janson L., Mackey L. Cross-validation Confidence Intervals for Test Error // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 16339–16350.
2. Beleites C., Baumgartner R., Bowman C., Somorjai R., Steiner G., Salzer R., Sowa M. G. Variance reduction in estimating classication error using sparse datasets // Chemometrics and Intelligent Laboratory Systems. 2005. Vol. 79, Iss. 1-2. P. 91—100. https://doi.org/10.1016/j.chemolab.2005.04.008
3. Franc V., Zien A., Schölkopf B. Support Vector Machines as Probabilistic Models // Proc. of the International Conference on Machine Learning (ICML). ACM, New York, USA, 2011. P. 665–672.
4. Friedman J., Hastie T., Tibshirani R. Additive logistic regression: a statistical view of boosting // Annals of Statistics. 2000. Vol. 28. P. 337–407. https://doi.org/10.1214/aos/1016218223
5. Кельманов А. В., Пяткин А. В. NP-трудность некоторых квадратичных евклидовых задач 2-кластеризации // Доклады Академии наук. 2015. Т. 464, № 5. С. 535–538. https://doi.org/10.7868/S0044466916030091
6. Лбов Г. С., Старцева Н. Г. Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. Новосибирск, 1990. Вып. 134 : Вычислительные системы. С. 56–66.
7. Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск : Институт математики СО РАН, 1999. 211 с.

8. Lugosi G., Vayatis N. On the bayes-risk consistency of regularized boosting methods // Annals of Statistics. 2004. Vol. 32. P. 30–55. https://doi.org/10.1214/aos/1079120129
9. Mease D., Wyner A. Evidence contrary to the statistical view of boosting // Journal of Machine Learning Research. 2008. Vol. 9. P. 131–156. https://doi.org/10.1145/1390681.1390687
10. Motrenko A., Strijov V., Weber G.-W. Sample Size Determination For Logistic Regression // Journal of Computational and Applied Mathematics. 2014. Vol. 255. P. 743–752. https://doi.org/10.1016/j.cam.2013.06.031
11. Krasotkina O. V., Turkov P. A., Mottl V. V. Bayesian Approach To The Pattern Recognition Problem In Nonstationary Environment // Lecture Notes in Computer Science. 2011. Vol. 6744. P. 24–29. https://doi.org/10.1007/978-3-642-21786-9_6.
12. Красоткина О. В., Турков П. А., Моттль В. В. Байесовская логистическая регрессия в задаче обучения распознаванию образов при смещении решающего правила // Известия Тульского государственного университета. Технические науки. 2013. № 2. С. 177–187.
13. Nedel'ko V. M. Misclassification probability estimations for linear decision functions. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2004. Vol. 3138. P. 780–787. https://doi.org/10.1007/978-3-540-27868-9_85
14. Nedel'ko V. Decision trees capacity and probability of misclassification // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNAI. 2005. Vol. 3505. P. 193–199. https://doi.org/10.1007/11492870_16
15. Неделько В. М. Регрессионные модели в задаче классификации // Сибирский журнал индустриальной математики. 2014. Т. 27, № 1. С. 86–98.
16. Неделько В. М. К вопросу об эффективности бустинга в задаче классификации // Вестник Новосибирского государственного университета. Серия: математика, механика, информатика. 2015. Т. 15, вып. 2. С. 72–89. https://doi.org/10.17377/PAM.2015.15.206
17. Torshin I. Yu., Rudakov K. V. On the Theoretical Basis of Metric Analysis of Poorly Formalized Problems of Recognition and Classification. Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications). 2015. Vol. 25, N 4. P. 577–587. https://doi.org/10.1134/S1054661815040252
18. Vanwinckelen G., Blockeel H. On estimating model accuracy with repeated cross-validation // BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning. 2012. P. 39–44.
19. Vorontsov K.V. Exact Combinatorial Bounds on the Probability of Overfitting for Empirical Risk Minimization // Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications). 2010. Vol. 20, N 3. P. 269–285. https://doi.org/10.1134/S105466181003003X

**Виктор Михайлович Неделько**, кандидат физико-математических наук, доцент, старший научный сотрудник, Институт математики им. С. Л. Соболева СО РАН, Российская Федерация, 630090, г. Новосибирск, просп. Академика Коптюга, 4, тел.: +7(383) 3332793, email: nedelko@math.nsc.ru, ORCID iD https://orcid.org/0000-0001-8217-9304

*Поступила в редакцию 30.10.2021*