АЛГЕБРО-ЛОГИЧЕСКИЕ МЕТОДЫ В ИНФОРМАТИКЕ
И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ALGEBRAIC AND LOGICAL METHODS IN COMPUTER
SCIENCE AND ARTIFICIAL INTELLIGENCE

Серия «Математика»
2020. Т. 33. С. 64—79

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

ИЗВЕСТИЯ
Иркутского
государственного
университета

УДК 519.246
MSC 68T10, 62H30
DOI https://doi.org/10.26516/1997-7670.2020.33.64

On Decompositions of Decision Function Quality Measure*

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

Abstract. A comparative analysis of two approaches to the decomposition of quality criterion of decision functions is carried out.

The first approach is the bias-variance decomposition. This is the most well-known decomposition that is used in analyzing the quality of decision function construction methods, in particular for justifying some ensemble methods. This usually assumes a monotonous dependence of the bias and variance on the complexity. Recent studies show that this is not always true.

The second approach (G.S. Lbov, N.G. Startseva, 1989) is a decomposition into a measure of adequacy and a measure of statistical stability (robustness). The idea of the approach is to decompose the prediction error into approximation error and statistical error.

In this paper we propose a method of statistical estimation of the components of both decompositions on real data. We compare the dependencies of these components on the

* This work was partly supported by the Russian Foundation for Basic Research, grants 18–07–00600, 19–29–01175 and by the Russian Science Foundation under grant 20-15-00057. The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no 0314-2019-0015).

complexity of the decision function. Non-normalized margin is used as a general measure of complexity.

The results of the study and the experiments on UCI data show significant qualitative similarities in behavior of the bias and the adequacy measure and between the variance and the statistical stability measure. At the same time, there is a fundamental difference between the considered decompositions, in particular, with increasing complexity, the measure of adequacy cannot increase, while the bias first decreases, but at high enough values of complexity usually starts to grow.

Keywords: machine learning, bias-variance decomposition, decision function complexity.

1. Introduction

In the field of machine learning, it quickly became clear that the key parameter on which the quality of a solution depends is complexity of class of decision functions [11] (it is more correct to speak of the complexity of the method for constructing decision functions) and dependence of quality on complexity has a specific form: first the error decreases, and then begins to grow [13]. To explain this phenomenon, error decompositions into components have been proposed.

The best known one is the bias-variance decomposition [20].

In the classical version of the decomposition, "variance" is understood as the average variance of predicted values, and the bias decomposition is the mean square of the difference between the average forecast and the optimal forecast. Averaging is performed over the features space and over samples (on which the decision function is constructed).

This decomposition is widely used both to explain the characteristic shape of the learning curve (the dependence of the quality of the solution on the method complexity) [9] and to justify ensemble methods [6; 8], such as a random forest. Basically examples of dependence of the expansion components on complexity has a typical qualitative form: the bias decreases monotonically, the variance increases monotonically, the sum of these quantities has a characteristic minimum. However, it was noticed that in practice, the bias at a sufficiently high complexity can also start to grow [4].

The decomposition into the measure of adequacy and measure of stability [12] is devoid of this feature. The idea behind the approach is to decompose the error into an approximation error and a statistical error.

More complex class of decision functions can potentially better approximate the optimal solution, but the actual accuracy may decrease due to statistical error when the solution is constructing by a sample.

2. The problem of constructing decision functions

To introduce the basic concepts, consider the general statement [14] of the problem of constructing decision functions [19; 21].

Let X be the space of values of the variables used for the forecast, and and Y be space of values of predicted variables, and let on a given σ-algebra of subsets of the set $D = X \times Y$ some probability measure is defined P.

In the future, we need to compute the mathematical expectations from various functions, so we introduce a notation

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

Generally X can be an arbitrary set with an arbitrary sigma algebra of events. However, in practice, we can assume that we are dealing with a system of random variables. In other words, the joint distribution function $F(x,y)$ is given, where x – vector composed of values from X. In this case, expectation of function $f(x,y)$ can be written as

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

The existence of a joint integral implies the existence of repeated integrals. With repeated integrals, you can enter the conditional distribution and the conditional expectation

$$\mathsf{E}_{X,Y} f(x,y) = \int_X \int_Y f(x,y)\, dF_{Y|x}(y)\, dF_X(x) = \int_X \mathsf{E}_{Y|x} f(x,y)\, dF_X(x).$$

The decision function is the correspondence $\lambda : X \to U$, i.e. a mapping from X to space of estimates U, which in a particular case may coincide with Y, and can, for example, represent estimates of the probabilities of an object belonging to given classes.

The quality of solution [16] is estimated by given loss function: $L : Y \times U \to [0,\infty)$.

By risk we mean average losses:

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

There are quality criteria [17], which are not expressed in terms of the integral of the loss function, for example, AUC. We do not consider them in this paper.

Mostly, the decision function is built on the basis of a random independent sample from the distribution P:

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

The sample size N will be fixed, so this parameter will be omitted in sample notation. The expectation of sample function is

$$\mathsf{E}_{D^N} f(V_N) = \int_{D^N} f(V_N) \ (\mathsf{P}(dx, dy))^N .$$

By the method of constructing decision functions we will call the mapping

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

where Λ is given class of decision functions, and W is set of all samples.

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

For a solution built on sample risk also becomes a function of sample:

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

Here we have introduced an abbreviation $R_Q(V_N)$ instead of $R(\lambda_{Q,V_N}, \mathsf{P})$ to emphasize that the criterion is a function of the sample, while the parameters Q and P are fixed. Since the expression is a function of sample, it is a random variable.

Quality of the method Q is characterized by the average risk

$$F_N(Q) = \mathsf{E}_{W_N} R_Q(V_N).$$

The problem of constructing decision functions (machine learning) is to find a method Q that would minimize $F_N(Q)$.

Note that in this statement there is no concept of "true dependence". The primary concept here is a criterion of quality that through the loss function reflects the objective losses by an erroneous decision. Moreover, this target criterion may not coincide with the empirical criterion, which is optimized for searching a solution [18].

To minimize $F_N(Q)$, an understanding of how a given value behaves is required. To get this, one uses decomposition of $F_N(Q)$ into components whose behavior is easier to analyze and explain.

The purpose of this work is to analyze the qualitative behavior of decompositions components.

3. Bias and variance

3.1. Decomposition idea

The idea of expansion is based on the following elementary formula, which is valid for any random variables ξ and ζ

$$\mathsf{E}(\xi-\zeta)^2=\mathsf{D}\xi+\mathsf{D}\zeta+(\mathsf{E}\xi-\mathsf{E}\zeta)^2-2\operatorname{cov}(\xi,\zeta).$$

To get the decomposition for the quadratic loss function, take y as ξ, and instead of ζ substitute $\lambda_{Q,V}(x)$. Note that these variables are independent. Mathematical expectations will be taken for the joint distribution on samples of volume N and for the random value y with given x:

We have:

$$F_N(Q)=\mathsf{E}_{W_N,X,Y}L(\lambda_{Q,V_N}(x),y)=\mathsf{E}_X\mathsf{E}_{W_N,Y|x}\left(\lambda_{Q,V_N}(x)-y\right)^2=$$
$$=\mathsf{E}_X\mathsf{D}_{Y|x}y+\mathsf{E}_X\mathsf{D}_{W_N}\lambda_{Q,V_N}(x)+\mathsf{E}_X\left(\mathsf{E}_{Y|x}y-\mathsf{E}_{W_N}\lambda_{Q,V_N}(x)\right)^2 \qquad (3.1)$$

The first term (the average conditional variance of y) is usually called "noise", the second term (the average conditional variance of the forecast) is the "variance", and the last term is the mean square of the bias.

3.2. Logarithmic loss function

Let us now consider the case of a logarithmic loss function. The corresponding decomposition was proposed in the work [5].

For two classes, the logarithmic loss function looks like

$$L(y,u_1)=-y\ln u_1-(1-y)\ln(1-u_1),$$

where u_1 is solution in the form of an estimate of the probability of $y=1$, i.e. the probability that the object belongs to the first class.

For the case of several classes, when $Y=\{1,\dots,k\}$, the solution is a vector $u=(u_1,...,u_k)$, $\sum\limits_{\omega=1}^{k}u_\omega=1$. Then the logarithmic loss function takes the form

$$L(y,u)=-\ln u_y,$$

where u_y is the solution component corresponding to class y.

In this form, the loss function is not suitable for decomposition, because instead of y, distribution cannot be substituted into it. However, the logarithmic loss function can be written in the form of the Kullback – Leibler divergence (we will denote it as $K(\cdot,\cdot)$).

Since divergence is only defined for distributions, we need to match the distribution to the Y variable. Binary vector can be used. We introduce random variables $(v_1,...,v_k)$, where v_ω is an indicator function for $y=\omega$.

We get

$$L(y,u) = -\sum_{\omega=1}^{k} v_\omega \ln u_\omega = \sum_{\omega=1}^{k} v_\omega \ln \frac{v_\omega}{u_\omega} = K(v,u).$$

In the last expression, we assume $0 \ln 0 = 0$, therefore $v_\omega \ln v_\omega \equiv 0$.

Due to independence of v and u we have

$$\mathsf{E}_{y,u} L(y,u) = -\sum_{\omega=1}^{k} \mathsf{E}_y v_\omega \mathsf{E}_u \ln u_\omega = -\sum_{\omega=1}^{k} p_\omega \mathsf{E}_u \ln u_\omega = \mathsf{E}_u K(p,u) + H(p),$$

where $p_\omega = \mathsf{E}_y v_\omega$ is the probability that $y = \omega$, and $H(p) = -\sum_{\omega=1}^{k} p_\omega \ln p_\omega$.

We introduce the vector $\bar{u} = (\bar{u}_1, ..., \bar{u}_k)$, whose components are computed as

$$\bar{u}_\omega = \frac{1}{C} e^{\mathsf{E}_u \ln u_\omega}, \quad C = \sum_{\omega=1}^{k} e^{\mathsf{E}_u \ln u_\omega},$$

where C is normalization constant that provides the condition $\sum_{\omega=1}^{k} \bar{u}_\omega = 1$. Actually $\bar{u}$ is just the normalized geometric mean of u vector .

Taking the logarithm, we get

$$\ln C = \mathsf{E}_u \ln u_\omega - \ln \bar{u}_\omega.$$

We multiply both sides of the equation by $\bar{u}_\omega$ and sum up over ω. We get

$$\ln C = \sum_{\omega=1}^{k} \bar{u}_\omega (\mathsf{E}_u \ln u_\omega - \ln \bar{u}_\omega) = -\mathsf{E}_u K(\bar{u}, u).$$

Similarly multiplying by p_ω and summing up, we have

$$\ln C = \sum_{\omega=1}^{k} p_\omega (\mathsf{E}_u \ln u_\omega - \ln \bar{u}_\omega) = K(p, \bar{u}) - \mathsf{E}_u K(p,u).$$

Equating the obtained expressions for $\ln C$, we obtain the required decomposition

$$\mathsf{E}_u K(p,u) = K(p,\bar{u}) + \mathsf{E}_u K(\bar{u},u).$$

Finally

$$\mathsf{E}_{y,u} L(y,u) = H(p) + K(p,\bar{u}) + \mathsf{E}_u K(\bar{u},u). \tag{3.2}$$

It is decomposition into noise, bias and variance.

3.3. General decomposition

Notice, that $\bar{u} = \arg\min_v \mathsf{E}_u K(v, u)$. Thus, the decomposition 3.2 fits into the general decomposition scheme [3].

Value $\bar{u}$ minimizing variation in [3] called the "main prediction".

In fact, this is not a "prediction", i.e. not the decision function value, but the distribution for the target variable at which the predictions made would have the least error. For a symmetric loss function, this remark is insignificant, but for an asymmetric one (such as divergence), it is necessary to correct the order of the arguments in the variance component from the general decomposition [3].

Note that the decomposition 3.2 would look more elegant with $K(\bar{u}, p)$ instead of $K(p, \bar{u})$. Indeed, the divergence is asymmetric and the first argument is the "true" distribution, and the second is its estimate.

For the two components of the decomposition, a general definition can be given that does not depend on a specific loss function.

Noise is the average loss for the optimal decision function.

The variance is the average loss at the "best" distribution for a given decision function method.

The bias can also be defined in a general way, if you define it simply as the remainder in the decomposition.

We will not separate the bias from the noise, since this is only possible on the model (when the distribution is known), but not on real data. Then

$$\mathsf{E}_{y,u} L(y, u) = \mathsf{E}_{y,u} L(y, \bar{u}) + \mathsf{E}_u K(\bar{u}, u). \tag{3.3}$$

In the last decomposition, all values may be estimated from the test sample.

4. The theory of statistical stability of decision functions

The best known is the error decomposition into bias and variance. However, probably even earlier, another decomposition was proposed [12]: to adequacy measure and stability.

The idea behind the approach is to decompose the error into an approximation error and a statistical error.

The more complex class of decision functions, the more accurately it can potentially approximate optimal solution, but the actual accuracy may decrease due to statistical error when solution is built on sample.

The basic concept of this decomposition is the asymptotic average risk or the asymptotic value of the average quality

$$F_\infty(Q) = \lim_{N \to \infty} F_N(Q). \tag{4.1}$$

The measure of adequacy is the difference between the asymptotic mean risk and the Bayesian risk. This measure shows how good a solution the method could give in the case of an unlimited sample (or when constructing solutions on the distributions themselves).

The measure of statistical stability is the difference between the average risk and the asymptotic one.

Note that it would be more correct to call the introduced components a measure of inadequacy and a measure of instability, since they characterize error rather than accuracy. To avoid terminological inconveniences, we will also use the terms approximation error and statistical estimation error.

The Bayesian level of error (risk) is exactly what in the expansion 3.1 called noise.

We see that the decompositions are similar: both have three components, one of which (noise) coincides.

Until now, we understood the method as an arbitrary mapping from samples to solutions.

With this definition, the method can change the algorithm for constructing the solution depending on the sample size. For example, it is quite possible to imagine such a method that, with a sample size of 100, builds a solution in the form of a tree, with a size of more than 100, a linear one, and with a size of more than 1000, uses a neural network.

In the described situation, the definition of the asymptotic average risk becomes incorrect. In fact, this definition implies a narrower concept of the method for constructing decision functions. Note that all actually used methods fit into this concept.

Almost all existing methods for constructing decision functions are methods that optimize an empirical criterion. Typically, the empirical criterion is the objective (target) loss function evaluated on sample plus a so-called regularizer. We assume that the regularizer form does not depend on the sample size.

Thus, a method is constructed as follows: an empirical quality criterion is set, an optimization problem is posed in a certain class of decision functions, and an algorithm for solving this problem is constructed.

Generally speaking, it has not been proven that in this case it is necessary to solve [7] optimization problem exactly. It can be assumed that an algorithm that finds a non-optimal solution from the point of view of an empirical criterion will find better solutions according the original (target) criterion. Moreover, the search approximation is a kind of regularization.

Decomposition into the measure of adequacy and stability requires narrowing the concept of the method for constructing decision functions, at the same time it is applicable to almost any quality criteria, for example, to AUC.

5. Bias and variance qualitative behavior

In the following sections, we examine the qualitative behavior of the decomposition components on real data. However, numerical experiments based on statistical modeling cannot be considered a rigorous proof of the assertions, therefore, we first consider synthetic examples in which the properties of decompositions can be determined analytically.

Example 1. Let the range of values of the target variable be a unit interval, i.e. $Y = [0, 1]$, and let there be a single variable x with a range $X = [0, 1]$.

Distribution on the set $X \times Y$ assume uniform.

Consider two methods for constructing decision functions. Both methods minimize the sample mean square loss function.

The first method uses decision functions of the form $u(x) = a$, where a is a constant selected from a sample, and the second method builds a solution in the form $u(x) = a + bx$, where a and b are constants tuned on (learned from) the sample, and $b \geqslant 0$.

Obviously, the complexity of the second method is higher than the first.

Proposition 3. *As the complexity of the method for constructing the decision function increases, the bias can increase.*

Proof. Let's estimate the biases for the methods from the example 1.

Due to the symmetry of the model, the average solution for the first method is $\bar{u}(x) \equiv 0{,}5$, whence we find that the bias is zero.

For the second method $\bar{u}(x)$ cannot be constant (due to $b \geqslant 0$, this function is increasing), so the bias is nonzero (since the optimal solution is a constant). □

The considered example includes a completely unnatural limitation $b \geqslant 0$. Such restrictions are hardly presented in the methods used in practice for constructing decision functions, however, the example is quite correct as evidence of the possibility of non-monotonic behavior of displacement.

Example 2. Let now $X = [0, 1]$, and $Y = (-\infty, +\infty)$.

Distribution on the set $X \times Y$ is set as follows: we put the distribution on X uniform, and the conditional distribution $\mathsf{P}(dy|x)$ let it be normal with parameters x and 1.

The decisive function will be constructed as follows.

Interval $X = [0, 1]$ is splitted into M equal parts, and then we shift all boundaries by the same value δ, which is chosen randomly (in accordance with a uniform distribution) from the interval $\left(0, \frac{1}{M}\right)$.

As a result, we get that X is divided into $M + 1$ parts, of which the first and the last have a random length, and the lengths of the rest partsare the same and equal $\frac{1}{M}$.

On each obtained segments, as a solution $u(x)$, we take the sample mean on this segment (if no points are included in the segment, then as a solution we assign the arithmetic mean of the coordinates of the ends of the segment).

Parameter M is characteristic of the complexity of this method.

Proposition 4. *As the complexity of the method for constructing the decision function increases, the variance can decrease.*

Proof. Let us estimate the dependence of the variance on M for the method from the example 2.

Due to the construction of the model, the mean solution $\bar{u}(x) = x$ on the interval $\left[\frac{1}{M}, 1 - \frac{1}{M}\right]$, hence the bias on this interval is equal to zero. On the edge segments, the bias does not exceed $\frac{1}{M^2}$. The average bias over the entire X can be estimated from above by the value $\frac{1}{M^3}$.

General error (mean square deviation) not less then $\frac{M-1}{M} \cdot \frac{1}{12M^2}$ and not more then $\frac{1}{12M^2}$

It follows, that the bias has lower estimate $\frac{M-1}{M} \cdot \frac{1}{12M^2} - \frac{1}{M^3}$ and upper estimate $\frac{1}{12M^2}$.

By direct substitution, we see that the lower estimate $M = 100$ is greater than the upper estimate for $M = 1000$. □

Remark 1. Defining the method for constructing decision functions as a function of sample, we did not provide possibility of using random parameters. However, this is insignificant, since we can always take a value composed of the least significant decimal places of the sample values as a random parameter. Formally, we will remain in the class of deterministic functions of sample, but in practice we will get a randomized solution.

6. Method of statistical modeling on real-world data set

Statistical modeling involves generating a large number of samples, therefore, as a rule, it is performed using artificial data set.

This paper proposes a statistical modeling method based on real–world data set. Now the volume of real data is often so large that you can work with it, in a sense, as with distributions. At the same time, the concept of "large" sample is obviously relative. You can use a sample as a distribution if its size is much larger than the size of the subsamples used for training.

The general scheme of modeling is as follows. The entire sample is divided into two approximately equal parts, which we will call the complete training sample and the test sample. The complete training sample is splited into a large number (of the order of tens or hundreds) of training samples.

We will use averaging over a large number of training samples as the mathematical expectation over the samples. We will interpret averaging over the test sample as the mathematical expectation over the distribution.

To estimate the limit 4.1 when the sample tends to infinity, we take the value when training on a full training set.

The described techniques are sufficient for estimating all components of the expansions, except for "noise", which we will not separate from the bias or from the measure of adequacy, respectively.

For further research, it is required to introduce a measure of complexity [1] for a method.

Different methods have different parameters that are responsible for the complexity of the solution. For example, complexity characteristics for tree ensembles are the number of trees and the depth of the tree. It is necessary to introduce a single characteristic of complexity.

Let's take a margin as a measure of complexity.

Recall that the margin can be normalized and non-normalized. [15].

We use non-normalized margin. This is the margin that ensembles monotonically increase in the process of learning (by construction). This is a natural measure of learning.

The normalized margin appears in the assessment of generalizing ability [10]. As a rule, this margin begins to decrease in the course of training.

Since not all implementations of algorithms for constructing decision functions provide the output of the margin value, we will calculate it through the probability estimates, namely

$$M(\lambda, V) = \sum_{i=1}^{N} y_i \ln \frac{\tilde{p}_i}{1 - \tilde{p}_i}. \tag{6.1}$$

This formula makes sense only for the case of two classes, when $Y = \{-1, 1\}$. Here $\tilde{p}_i$ denotes an estimate of the probability of an object belonging to class 1.

This expression reflects a real connection [15] between margin and probability estimates for methods that support the concept of margin (logistic regression, boosting). For the rest of the methods we will consider 6.1 as the definition of the average (for the sample) margin.

We can only estimate the noise approximately. But we do not need it at all, since in every task it is a constant.

7. Experimental results

The Adult Data Set (UCI Machine Learning Repository) was taken as data for experimental research. The number of objects is 50,000; for simplicity, 8 numeric variables have been selected.

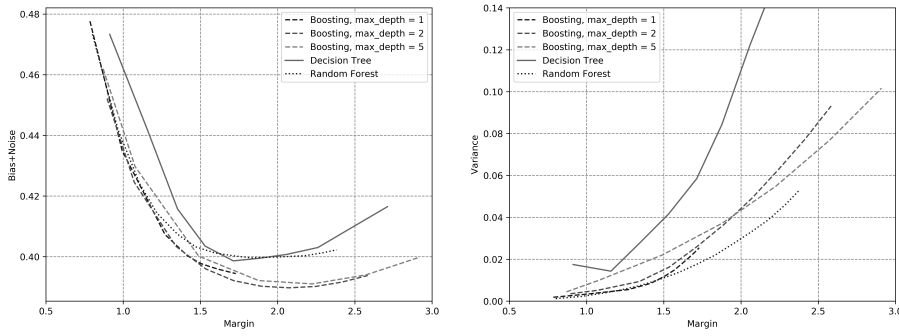

Figure 1. Dependencies of bias and variance on margin (complexity parameter) for various classification methods.

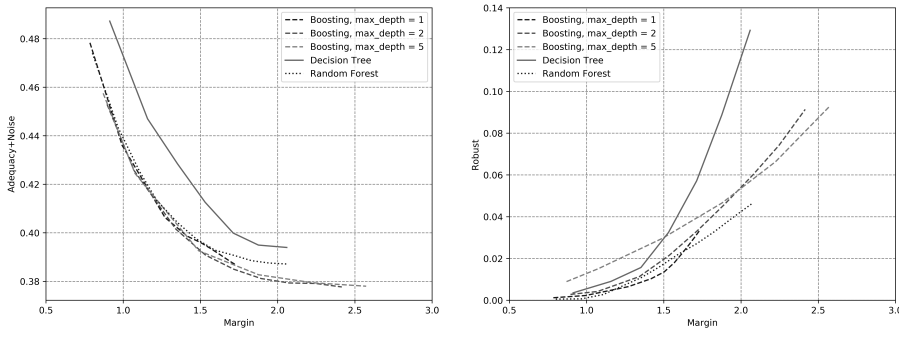

Figure 2. Dependencies of the measure of adequacy and measure of stability on the margin (complexity parameter) for various classification methods.

The following classification methods [2] based on trees were taken: decision tree, random forest and gradient boosting. For the decision tree and for the random forest, the margin (complexity) was changed by changing the maximum tree depth. The number of trees in the ensemble of a random forest has practically no effect on the complexity, so it was assigned at 100. For boosting, the complexity depends on both the depth and the number of trees, so three curves were built for fixed depth values 1, 2, and 5 with a varying number of trees. We see (Fig. 1)that the qualitative form of the dependencies is similar, while the bias does not always decrease monotonically, but starts to grow from some moment.

In fig. 2 for the same methods, the dependencies for the measure of adequacy and the measure of stability are shown. These components are monotonic due to their properties. Despite the qualitative difference associated with monotonicity, both decompositions are quite similar.

8. Conclusion

The paper proposes a statistical modeling scheme that allows us to estimate the decompositions components on real–world datasets.

A general measure of complexity is proposed: non-normalized margin.

Comparison of two decompositions was carried out: for bias and variance and for measure of adequacy and stability. It is shown that with increasing complexity, the bias can increase, and the variance can decrease, while the decomposition into the measure of adequacy and stability always has a "canonical" form.

References

1. Baliuk A.S. Complexity Lower Bound for Boolean Functions in the Class of Extended Operator Forms. *The Bulletin of Irkutsk State University. Series Mathematics*, 2019, vol. 30, pp. 125-140. https://doi.org/10.26516/1997-7670.2019.30.125
2. Berikov V. Semi-Supervised Classification Using Multiple Clustering And Low-Rank Matrix Operations. *Lecture Notes in Computer Science*, 2019, vol. 11548, pp. 529-540.
3. Domingos P. A Unified Bias-Variance Decomposition and its Applications. *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, Stanford, CA: Morgan Kaufmann, pp. 231-238.
4. Dyakonov A.G. Bias and variance. *https://dyakonov.org/page/2/*, 2016. (In Russian)
5. Heskes T. Bias/Variance Decompositions for Likelihood-Based Estimators. *Neural Computation*, 1998, vol. 10, no. 6, pp. 1425–1433.
6. Kanevskiy D., Vorontsov K. Cooperative Coevolutionary Ensemble Learning. *Multiple Classifier Systems, 7th International Workshop, MCS 2007*, Prague, Czech Republic, May 23–25. 2007. Proceedings, pp. 469-478. https://doi.org/10.1007/978-3-540-72523-7_47.
7. Khachay M., Khachay D. Attainable accuracy guarantee for the k-medians clustering in [0, 1]. *Optimization Letters*, 2019, vol. 13, no. 8, pp. 1837-1853. https://doi.org/10.1007/s11590-018-1305-3
8. Kotsiantis S. Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review*, 2014, no. 29(1), pp. 78-100. https://doi.org/10.1017/S0269888913000313
9. Kuncheva L.I., Skurichina M., Duin R.P.W. An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion*, 2002, no. 3, pp. 245-258.
10. Kuncheva L., Vetrov D. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, vol. 28, no. 11, pp. 1798-1808.
11. Lbov G.S., Startseva N.G. On Some a Concept of Complexity of a Strategy of Nature in Pattern Recognition. *Data Analysis in Expert Systems. Computational systems*, Novosibirsk, 1986, issue 117, pp. 91-102. (In Russian).
12. Lbov G.S., Startseva N.G. Complexity of Distributions in the Classification Problem. *Doklady RAS*, 1994, vol. 338, no. 5, pp. 592-594. (In Russian).
13. Lbov G.S., Startseva N.G. *Logicheskie reshajushhie funkcii i voprosy statisticheskoj ustojchivosti reshenij* [Logical Decision Functions and the Problem of

Statistical Robustness of Solutions]. Novosibirsk, Institute of Mathematics Publ., 211 p. (In Russian).
14. Nedel'ko V.M. Some aspects of estimating a quality of decision functions construction methods. *Tomsk State University Journal of Control and Computer Science*, 2013, no. 3 (24), pp. 123-132. (In Russian).
15. Nedel'ko V.M. On performance of boosting in classification problem. *Vestn. Novosib. Gos. Univ., Ser. Mat. Mekh. Inform*, 2015, vol. 15, no. 2, pp. 72-89. (In Russian).
16. Nedel'ko V.M. Tight risk bounds for histogram classifier. *Proceedings of IFOST-2016 11th International Forum on Strategic Technology*, 2016, pp. 267-271.
17. Nedel'ko V.M. On the Maximization of Quadratic Weighted Kappa. *The Bulletin of Irkutsk State University. Series Mathematics*, 2018, vol. 23, pp. 36-45. https://doi.org/10.26516/1997-7670.2018.23.36 (In Russian).
18. Nedel'ko V.M. Statistical Fitting Criterion on the Basis of Cross-Validation Estimation. *Pattern Recognition and Image Analysis*, ISSN 1054-6618, Pleiades Publishing, Ltd., 2018, vol. 28, no. 3, pp. 510-515. https://doi.org/10.1134/S1054661818030148
19. Rudakov K.V. Mathematical Foundations for Processing High Data Volume, Machine Learning, and Artificial Intelligence. *Pattern Recognit. Image Anal*, 2019, vol. 29, pp. 339-343. https://doi.org/10.1134/S1054661819030192
20. Stuart G., Bienenstock E., Doursat R. Neural networks and the bias/variance dilemma. *Neural Computation*, 1992, vol. 4. DOI: 10.1162/neco.1992.4.1.1.
21. Zhuravlev Y.I., Ryazanov V.V., Aslanyan L.H. et al. On a Classification Method for a Large Number of Classes. *Pattern Recognit. Image Anal*, 2019, vol. 29, pp. 366-376. https://doi.org/10.1134/S1054661819030246

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

Received 03.06.2020

О разложениях критериев качества решающих функций

В. М. Неделько

Институт математики им. С. Л. Соболева, Новосибирск, Российская Федерация

Аннотация. Проводится сравнительный анализ двух подходов к разложению критерия качества решающих функций. Первый подход — разложение на смещение и разброс (bias-variance decomposition). Второй подход — разложение на меру адекватности и меру статистической устойчивости. Выявлено качественное сходство получаемых разложений, при этом установлено (как аналитически, так и на основе численного эксперимента), что только второй подход гарантирует монотонность компонент разложения.

Ключевые слова: построение решающих функций, распознавание образов, bias-variance decomposition, статистическая устойчивость.

Список литературы

1. Baliuk A. S. Complexity Lower Bound for Boolean Functions in the Class of Extended Operator Forms // Известия Иркутского государственного университета. Серия Математика. 2019. Т. 30. С. 125–140. https://doi.org/10.26516/1997-7670.2019.30.125
2. Berikov V. Semi-Supervised Classification Using Multiple Clustering And Low-Rank Matrix Operations // Lecture Notes in Computer Science. 2019. Vol. 11548 LNCS. P. 529–540.
3. Domingos P. A Unified Bias-Variance Decomposition and its Applications // Proceedings of the Seventeenth International Conference on Machine Learning. Stanford, CA: Morgan Kaufmann, 2000. P. 231–238.
4. Дьяконов А. Г. Смещение (bias) и разброс (variance). URL: https://dyakonov.org/page/2/
5. Heskes T. Bias/Variance Decompositions for Likelihood-Based Estimators // Neural Computation. 1998. Vol. 10, N 6. P. 1425–1433.
6. Kanevskiy D., Vorontsov K. Cooperative Coevolutionary Ensemble Learning // Multiple Classifier Systems, 7th International Workshop, MCS 2007, Prague, Czech Republic, May 23–25. 2007. Proceedings. P. 469–478. https://doi.org/10.1007/978-3-540-72523-7_47.
7. Khachay M., Khachay D. Attainable accuracy guarantee for the k-medians clustering in [0, 1] // Optimization Letters. 2019. Vol. 13, N 8. P. 1837–1853. https://doi.org/10.1007/s11590-018-1305-3
8. Kotsiantis S. Bagging and boosting variants for handling classifications problems: A survey // The Knowledge Engineering Review. 2014. N 29(1). P. 78–100. https://doi.org/10.1017/S0269888913000313
9. Kuncheva L. I., Skurichina M., Duin R. P. W. An experimental study on diversity for bagging and boosting with linear classifiers // Information Fusion. 2002. N 3. P. 245–258.
10. Kuncheva L., Vetrov D. Evaluation of stability of k-means cluster ensembles with respect to random initialization // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006. Vol. 28, N 11. P. 1798–1808.
11. Лбов Г. С., Старцева Н. Г. Об одном понятии сложности стратегии природы в распознавании образов // Анализ данных в экспертных системах. Новосибирск, 1986. Вып. 117 : Вычислительные системы. С. 91–102.
12. Лбов Г. С., Старцева Н. Г. Сложность распределений в задачах классификации // Доклады РАН. 1994. Т. 338, № 5. С. 592–594.
13. Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск : Издательство Института математики, 1999. 211 с.
14. Неделько В. М. Некоторые вопросы оценивания качества методов построения решающих функций // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 3 (24). С. 123–132.
15. Неделько В. М. К вопросу об эффективности бустинга в задаче классификации // Сибирский журнал чистой и прикладной математики. 2015. Т. 15, вып. 2. С. 72–89.
16. Nedel'ko V. M. Tight risk bounds for histogram classifier // Proceedings of IFOST-2016 11th International Forum on Strategic Technology IFOST-2016. 2016. P. 267–271.

17. Неделько В. М. О максимизации критерия квадратичного взвешенного каппа // Известия Иркутского государственного университета. Серия Математика. 2018. Т. 23. С. 36–45. https://doi.org/10.26516/1997-7670.2018.23.36
18. Nedel'ko V. M. Statistical Fitting Criterion on the Basis of Cross-Validation Estimation // Pattern Recognition and Image Analysis. 2018. Vol. 28, N 3. P. 510–515. https://doi.org/10.1134/S1054661818030148
19. Rudakov K. V. Mathematical Foundations for Processing High Data Volume, Machine Learning, and Artificial Intelligence // Pattern Recognit. Image Anal. 2019. Vol. 29. P. 339–343. https://doi.org/10.1134/S1054661819030192
20. Stuart G., Bienenstock E., Doursat R. Neural networks and the bias/variance dilemma // Neural Computation. 1992. Vol. 4. https://doi.org/10.1162/neco.1992.4.1.1.
21. Zhuravlev Y. I., Ryazanov V. V., Aslanyan L. H. et al. On a Classification Method for a Large Number of Classes // Pattern Recognit. Image Anal. 2019. Vol. 29. P. 366–376 . https://doi.org/10.1134/S1054661819030246

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

Поступила в редакцию 03.06.2020