ISSN 2304-0122 Ufa Mathematical Journal. Vol. 10. No 1 (2018). P. 49-63.

doi:10.13108/2018-10-1-49 UDC 519.25

COMBINATORIAL BOUNDS OF OVERFITTING FOR THRESHOLD CLASSIFIERS

Sh.Kh. ISHKINA

Abstract. Estimating the generalization ability is a fundamental objective of statistical learning theory. However, accurate and computationally efficient bounds are still unknown even for many very simple cases. In this paper, we study one-dimensional threshold decision rules. We use the combinatorial theory of overfitting based on a single probabilistic assumption that all partitions of a set of objects into an observed training sample and a hidden test sample are of equal probability. We propose a polynomial algorithm for computing both probability of overfitting and of complete cross-validation. The algorithm exploits the recurrent calculation of the number of admissible paths while walking over a three-dimensional lattice between two prescribed points with restrictions of special form. We compare the obtain sharp estimate of the generalized ability and demonstrate that the known upper bound are too overstated and they can not be applied for practical problems.

Keywords: computational learning theory, empirical risk minimization, combinatorial theory of overfitting, probability of overfitting, complete cross-validation, generalization ability, threshold classifier, computational complexity.

Mathematics Subject Classification: 68Q32, 60C05

1. Introduction

We consider the following mathematical model of decisions making under the incompleteness of an information. We are given a binary matrix and its rows correspond to objects and the columns correspond to the rules of decisions making called also classifiers or predictors. An entry of the matrix is one if and only if a given classifier makes an error at the given object. In the set $\mathbb{X}$ of all rows of the matrix, we choose randomly and equiprobably the observed training sample, a subset $X \subset \mathbb{X}$ of a fixed cardinality. Then in the set $\mathbb{A}$ of all columns of the matrix we choose the classifier with the minimal error rate on X. We want to estimate the the error rate of this classifier on a hidden testing sample $\bar{X} = \mathbb{X} \backslash X$. If the difference of the error frequencies on the testing and training samples exceeds ε, one says that an overfitting occurs. The obtaining of upper bounds for the probability of overfitting is one of the main issues in the statistical learning theory [1]–[3].

Classical Vapnik–Červonenkis estimates [1] depend only on the size of the error matrix. Being the "worst case" estimates, they are overstated by orders and do not fit well the experimental results [4]. More gentle estimates depend on the properties of partial order relations on the set of vector columns of the error matrix [5]. In the combinatorial theory of overfitting [6]–[8], there was justified the necessity of combination of two properties, the splitting and connectivity [9, 12]. Thanks to the splitting, the classifiers with a high probability of error make a negligibly small contribution into the overfitting. Thanks to the connectivity, the contribution into the overfitting by the classifiers with close error vectors reduces essentially.

Sh.Kh. Ishkina, Combinatorial bounds of overfitting for threshold classifiers.
The work is supported by RFBR under projects no. 15-37-50350 mol_nr and no. 14-07-00847.

Submitted December 21, 2016.

In [13] there were obtained conditions, under which splitting and connectivity bounds *were sharp*. In particular, they are satisfied by monotone and unimodal chains of classifiers [9]. In practical problems of statistical learning such chains can be generated by elementary threshold classifiers used in such classification algorithms as decision trees, logic algorithms [14], algorithms for calculating estimates [15], as well as in constructing linear classifiers by the method of coordinate-wise optimization. At that, one usually assumes the existence of error-free classifier that is almost impossible in practical problems. In the general case, the threshold classifiers generate sequences of classifiers called direct chains. Earlier, for them, there were known only the upper bound of the expected error rate on the testing sample [16]. Various specifications of the splitting and connectivity bounds taking into consideration, for instance, the pairwise competition between the classifiers [17] or the stratified clustering of the set of classifiers [18, 19] are still overstated for the direct chains.

In the present work we propose an algorithm of polynomial complexity for calculating the probability of overfitting for an arbitrary direct chain. The algorithm is based on the recurrent calculation of the number of admissible paths while walking over a three-dimensional set between two given points with restrictions of special form.

1.1. Main definitions. We are given a finite set $\mathbb{X} = \{x_1, \ldots, x_L\}$, whose elements are called *objects* and a finite set $\mathbb{A}$, whose elements are called *classifiers*. The set $\mathbb{A}$ is called the *family of classifiers*.

We are given a function $I\colon \mathbb{A} \times \mathbb{X} \to \{0, 1\}$ called *indicator function*. If $I(a, x) = 1$, we say that the classifier a *makes an error* at the object x. The binary matrix $\big(I(a, x)\colon x \in \mathbb{X}, a \in \mathbb{A}\big)$ of the size $|\mathbb{X}|\times|\mathbb{A}|$ is called *error matrix*.

We suppose that each classifier $a \in \mathbb{A}$ is in one-to-one correspondence with its error vector $(I(a, x_i))_{i=1}^{L}$, that is, the error matrix can not contain two equal columns. We assume that the row ordering in the error matrix is not important. By a we shall denote both the classifier and its error vector.

A *number of errors* of a classifier a on a sample $X \subset \mathbb{X}$ is the quantity

$$n(a, X) = \sum_{x \in X} I(a, x).$$

An *error rate* of a classifier a on a sample $X \subset \mathbb{X}$ is the quantity

$$\nu(a, X) = n(a, X)/|X|.$$

By $[X]^l$ we denote the set of all subsets $\mathbb{X}$ of a cardinality $l < L$. The subsets $X \in [X]^l$ are called *training samples*, and their complements $\bar{X} = \mathbb{X}\backslash X$ are called *testing samples*. On the set $[X]^l$ we introduce the uniform probability distribution:

$$\mathsf{P}(X) = 1/C_L^l, \qquad X \in [X]^l.$$

An *overfitting* of a classifier a on a partition $(X, \bar{X})$ is the quantity

$$\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X).$$

If $\delta(a, X) > \varepsilon$, we say that the classifier a is overfitted on X.

A *learning algorithm* is the mapping $\mu\colon [X]^l \to \mathbb{A}$, which to each training sample X, a classifier $a = \mu X$ in the family $\mathbb{A}$ associates to.

A *pessimistic empirical risk minimization (PERM)* is a learning algorithm, which chooses a classifier making the smallest number of errors on a training sample X. If there are several such classifiers in the family, then it chooses the classifier with the maximal number of error on a testing sample $\bar{X}$ [9].

For a given learning algorithm μ, a family of classifiers $\mathbb{A}$, a set $\mathbb{X}$ and a size l of a training sample, a *probability of overfitting* is the functional

$$Q_\varepsilon(\mu, \mathbb{A}, \mathbb{X}, l) = \mathsf{P}[\delta(\mu X, X) \geqslant \varepsilon] = \frac{1}{C_L^l} \sum_{X \in [X]^l} [\delta(\mu X, X) \geqslant \varepsilon].$$

Hereinafter the square brackets stand for transformation of a logical condition into the numerical value by the rule [true] = 1, [false] = 0.

A *complete cross-validation* is the functional equal to the expectation of the number of the errors on a testing sample:

$$CCV(\mu, \mathbb{A}, \mathbb{X}, l) = \mathsf{E}\nu(\mu X, \bar{X}) = \frac{1}{C_L^l} \sum_{X \in [X]^l} \nu(\mu X, \bar{X}).$$

An effective calculation of Q_ε and CCV directly by definition is possible only for small $|\bar{X}| = L - l$. If l is close to $L/2$, the number of terms is exponentially large in L.

1.2. Direct sequences of classifiers. We consider the sets of objects, by which neighbouring classifiers in the set $\mathbb{A} = \{a_0, \ldots, a_P\}$ differ:

$$G_p = \big\{x \in \mathbb{X} \,|\, I(a_p, x) \neq I(a_{p+1}, x)\big\}, \quad p = 0, \ldots, P-1. \tag{1}$$

Definition 1. *The set of classifiers is called a direct sequence if the sets G_p are mutually disjoint.*

We note that it follows from the definition that the order of the classifiers is important. Indeed, we consider two families of the classifiers; the first being a direct sequence $\mathbb{A} = \{a_0, \ldots, a_P\}$, while the other is obtained from the first one by a permutation of the classifiers a_p and a_{p+1} for some p: $\mathbb{A}' = \{a_0, \ldots, a_{p-1}, a_{p+1}, a_p, a_{p+2}, \ldots, a_P\}$.

We define the sets G_p by (1). Then the family $\mathbb{A}'$ is not a direct chain since the neighbouring classifiers a_{p-1} and a_{p+1} differ by the set of objects $G_{p-1} \sqcup G_p$, while the classifiers a_{p+1} and a_p differ by the set of objects G_p, that is, these sets intersect.

Definition 2. *A direct sequence $\mathbb{A} = \{a_0, \ldots, a_P\}$ is called a direct chain if each pair of neighbouring classifiers differs by one object: $|G_p| = 1$, $p = 0, \ldots, P-1$. The number P is called a length of the direct chain $\mathbb{A}$.*

Definition 3. *A one-dimensional threshold classifier over a set $\mathbb{X} \subset \mathbb{R}$ is the family of threshold rules $a(x, \theta) = [x \geqslant \theta]$, where $\theta \in \mathbb{R}$ is a parameter called threshold.*

According to the following theorem, the notion of the direct sequence and the one-dimensional threshold classifier are synonyms.

Theorem 1. *We define the set V of direct sequences $\mathbb{A} = \{a_0, \ldots, a_P\}$ such that $\sum_{p=0}^{P-1} |G_p| = L$, where G_p are defined by (1) and the set U of one-dimensional classifiers over the set $\mathbb{X} = \{x_1, \ldots, x_L\}$ of the points in the real axis such that for each x_i, the true mark in the class $y_i \in \{0, 1\}$ corresponds to. Then there exists a bijection between these sets.*

Proof. In the sets V and U the objects are defined up to renaming the objects in the set $\mathbb{X}$.

Each object $u \in U$ is uniquely determined by the location of objects in two classes $\{0, 1\}$ on the real axis, that is, by the location of the points in the set $\mathbb{X}$ on the axis $\mathbb{R}$ and by the set of correct answers $\{y_1, \ldots, y_L\}$. The values of the thresholds are chosen so that they partition the set $\mathbb{X}$ into two classes in all possible ways.

Each object in the set V is uniquely determined by the numbers of units in the vector a_0, that is, $n(a_0, \mathbb{X})$, and by the sequence of pairs $(n_0^p, n_1^p)_{p=0}^{P-1}$, where n_0^p is the number of zeroes in the vector a_p being the units in the a_{p+1}, and n_1^p is the number of units in the vector a_p being the zeroes in a_{p+1}. Under the presence of such information, the error matrix $\{a_0, \ldots, a_P\}$

is constructed as follows. The vector a_0 is defined so that at the first $n(a_0, \mathbb{X})$ positions this vector contains the units followed by the zeroes. For each p, consequently starting from $p = 0$, the vector a_{p+1} is obtained from the vector a_p by inverting n_0^p zeroes and n_1^p units.

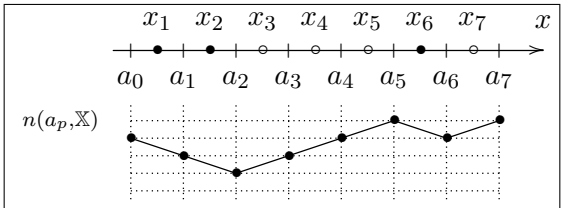
We construct a mapping $f : U \to V$ as follows. Assume we are given an object $u \in U$, that is, the set of the points $x_1 \leqslant \cdots \leqslant x_L$ and of correct answers $y_1, \ldots, y_L$. To this object, we associate a direct sequence $v = f(u) \in V$.

In order to do this, we introduce the indicator function $I(a, x_i) = \big[a(x_i, \theta) \neq y_i\big]$. Variation of θ generates at most $L + 1$ classifiers with pairwise distinct error vectors. They form a direct sequence. If all objects x_i are pairwise distinct, $x_1 < x_2 < \cdots < x_L$, then the direct sequence is a direct chain.

The mapping f determines uniquely a direct chain by the family of threshold rules, that is, it is an injection. Let us prove that it is a surjection.

Assume we are given a direct sequence $v \in V$, that is, the quantity $n(a_0, \mathbb{X})$ and the set of pairs $(n_0^p, n_1^p)_{p=0}^{P-1}$. Let us construct the error matrix $\{a_0, \ldots, a_P\}$. We define the family of threshold rules $u \in U$ as follows. To each set G_p, we associate the points $x_p^1 = \cdots = x_p^{|G_p|}$ and we let $x_0^1 < x_1^1 < \cdots < x_{P-1}^1$. We let $y_p^i = 1$ if $I(a_p, x_p^i) = 0$ and $y_p^i = 0$ otherwise. It is easy to check that the constructed family u is the pre-image of v under the mapping f, that is, $v = f(u)$. Thus, the mapping f is a bijection. □

Example 1. *In figure 1, an example of a direct chain is shown. At the axis x we indicate the objects x_i. The correct answers y_i are shown by the points ∘ and •. The thresholds θ are chosen in the centers between the neighbouring objects. Below we show the graph of the number of errors of classifiers and the error matrix.*



x_1	1	0	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0	0
x_3	0	0	0	1	1	1	1	1
x_4	0	0	0	0	1	1	1	1
x_5	0	0	0	0	0	1	1	1
x_6	1	1	1	1	1	1	0	0
x_7	0	0	0	0	0	0	0	1

FIGURE 1. Example of a direct chain

Definition 4. *A direct chain $\mathbb{A} = \{a_0, \ldots, a_P\}$ is called increasing (decreasing) if each classifier a_p makes an error $m + p$ times (respectively, $m - p$ times) on the set $\mathbb{X}$ for some value m. A direct chain $\mathbb{A}$ is called monotone if it is decreasing or increasing.*

A direct chain $\mathbb{A}$ can consist of several monotonicity segments. For instance, the chain shown in figure 1 contains for monotonicity segments: $\{a_0, a_1, a_2\}$ and $\{a_5, a_6\}$ are decreasing, $\{a_2, a_3, a_4, a_5\}$ and $\{a_6, a_7\}$ are increasing.

1.3. Formulation of problem. We want to find a way for calculating the functionals of the overfitting probability Q_ε and of the complete cross-validation CCV for PERM μ and for an arbitrary direct sequence $\mathbb{A}$ in a polynomial in L time.

2. Overfitting of an arbitrary family

We are given an arbitrary subset $\mathbb{D} \subseteq \mathbb{X}$ of the set $\mathbb{X}$. Each partition $(X, \bar{X})$ of the set $\mathbb{X} = X \sqcup \bar{X}$ induces the partition $(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D})$ of the subset $\mathbb{D}$. Also each pair of partitions $(D', \bar{D}')$ and $(D'', \bar{D}'')$ of subsets $\mathbb{D}' \subseteq \mathbb{X}$ and $\mathbb{D}'' = \mathbb{X} \backslash \mathbb{D}'$, respectively, defines a partition $(X, \bar{X})$ of the set $\mathbb{X}$ by the rule $X = D' \cup D''$ and $\bar{X} = \bar{D}' \cup \bar{D}''$.

A pair of classifiers a and a' are called *indiscernible on a set* $\mathbb{X}' \subset \mathbb{X}$ if $I(a, x) = I(a', x)$ for all $x \in \mathbb{X}'$.

Assume that we are given an arbitrary family of classifiers $\mathbb{A}$ and on the set $\mathbb{A} \times \mathbb{A} \times [X]^\ell$ we are given a strict order relation $a \succ_X a'$. We call it *finite* if for all classifiers $a, a' \in \mathbb{A}$ indiscernible on a set $\mathbb{X}' \subset \mathbb{X}$, the relation $a \succ_X a'$ is independent of the choice of the partition of the set $\mathbb{X}'$.

Example 2. *The order relations defined by the rules*

1. $a \succ_X a' \iff n(a, X) < n(a', X)$;
2. $a \succ_X a' \iff \delta(a, X) > \delta(a', X)$;

are finite.

Indeed, for each $X \in [X]^\ell$ and for each $\mathbb{X}'$, the identity

$$n(a, X) = n(a, X \cap \mathbb{X}') + n(a, X \setminus \mathbb{X}')$$

holds. If the classifiers a and a' are indiscernible on the set $\mathbb{X}'$, then $n(a, \mathbb{X}' \cap X) = n(a', \mathbb{X}' \cap X)$. This implies the finiteness of relation 1.

To prove the finiteness of relation 2, we rewrite the overfitting as

$$\delta(a, X) = \frac{1}{L - \ell} n(a, \mathbb{X}) - \frac{L}{(L - \ell)\ell} n(a, X).$$

Then the desired property follows the first statement.

The definition implies the following property:

Lemma 1. *Assume that the classifiers of the family $\mathbb{A}' \subseteq \mathbb{A}$ are indiscernible on the set $\mathbb{N}'$. Then for each $a \in \mathbb{A}'$, the validity of a finite relation $a \succ_X a'$ simultaneously for all $a' \in \mathbb{A}' \setminus \{a\}$ is independent of the choice of the partition of the set $\mathbb{N}'$.*

We say that on a sample X a classifier a is *better than* a' if $a \succ_X a'$. We call a learning algorithm $\mu\colon [X]^\ell \to \mathbb{A}$ *finite* if the learning result is the classifier best from the point of view of a finite relation $\succ_X$:

$$a = \mu X \quad \Leftrightarrow \quad a \succ_X a', \quad \forall a' \neq a. \tag{2}$$

Example 3. *The algorithm of empirical risk minimization (ERM) choosing the classifier with the minimal number of errors on a training sample and the algorithm of overfitting maximization (OM) choosing the classifier with the maximal overfitting are finite.*

The OM method arises in the problem of combinatorial calculation of Rademacher complexity of the class of decision rules [10]. *Indeed, as $\ell = \frac{L}{2}$, the random variables*

$$\sigma_i = \begin{cases} +1 & \text{if} \quad x_i \in \bar{X}, \\ -1 & \text{if} \quad x_i \notin \bar{X} \end{cases}$$

obey the Rademacher distribution $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$. Then the Rademacher complexity of the family is equal to the expectation of the overfitting of the OM method μ [11]*:*

$$\mathcal{R}_L(\mathbb{A}, \mathbb{X}) = \mathsf{E} \sup_{a \in \mathbb{A}} \frac{2}{L} \sum_{i=1}^{L} \sigma_i a_i = \mathsf{E} \sup_{a \in \mathbb{A}} \nu(a, \bar{X}) - \nu(a, X) = \mathsf{E}\delta(\mu, \bar{X}).$$

The Rademacher complexity can be considered as a quantity describing the complexity of the class of decision rules. The more the Rademacher complexity, the errors of classifiers can correlate better with the random noise σ_i.

By $\mathbb{D}$ we denote the subset of objects by which the classifiers of the family $\mathbb{A} = \{a_0, \ldots, a_P\}$ are discernible

$$\mathbb{D} = G_0 \cup \cdots \cup G_{P-1} = \big\{x \in \mathbb{X} \,|\, \exists\, a, a' \in \mathbb{A}\colon I(a,x) \neq I(a',x)\big\}, \tag{3}$$

where the sets G_p are defined according to (1).

We call the objects in the set $\mathbb{N} = \mathbb{X} \setminus \mathbb{D}$ *neutral*. On the set $\mathbb{N}$ the classifiers of the family are indiscernible and make the same number m of errors. By m_p we denote the number of errors of the classifier a_p on the set $\mathbb{D}$:

$$\begin{aligned} &m = n(a, \mathbb{N}), \quad \forall a \in \mathbb{A}; \\ &m_p = n(a_p, \mathbb{D}). \end{aligned} \tag{4}$$

Let us reduce the problem on calculating the probability of ovefitting Q_ε and of the complete cross-validation CCV to finding the number of partitions of the set $\mathbb{D}$ with some restrictions.

We denote by t the number of the objects in $\mathbb{D}$ contained in the training sample X, while by e we denote the number of errors of a classifier a_p on these objects. We introduce two functions of t and e: the number of partitions of the set $\mathbb{N}$ such that the classifier a_p is overfitted on X

$$N_p(t,e) = \#\big\{(X \cap \mathbb{N}, \bar{X} \cap \mathbb{N}) \;\big|\; \delta(a_p, X) \geqslant \varepsilon,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\},$$

and the number of the partitions of the set $\mathbb{D}$ such that a_p is the result of the learning:

$$D_p(t,e) = \#\big\{(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D}) \;\big|\; \mu X = a_p,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\}.$$

We introduce a *hypergeometric distribution function*

$$H_L^{l,m}(s) = \frac{1}{C_L^l} \sum_{i=0}^{\min\{\lfloor s \rfloor, l, m\}} C_m^i C_{L-m}^{l-i},$$

where $\lfloor x \rfloor$ is the integer part of x, that is, the greatest integer not exceeding x. Given a set $\mathbb{X}$ of a cardinality L and a sample $X_0 \subset \mathbb{X}$ of size m, the hypergeometric distribution function $H_L^{l,m}(s)$ is equal to the part of the sample of the set $\mathbb{X}$ of size l containing at most s elements in X_0. We let $C_n^i = 0$ if the condition $0 \leqslant i \leqslant n$ fails.

Theorem 2. *Given an arbitrary family of classifiers $\mathbb{A} = \{a_0, \ldots, a_P\}$, a finite learning algorithm μ, a set $\mathbb{X}$ of a cardinality L, a size l of a training sample, a precision $\varepsilon \in (0,1)$, the probability of overfitting is of the form*

$$Q_\varepsilon = \frac{1}{C_L^l} \sum_{p=0}^{P} \sum_{(t,e) \in \Psi_p} D_p(t,e) N_p(t,e), \tag{5}$$

where the set $\mathbb{D}$, the parameters m_p and m are determined by (3) *and* (4) *and*

$$\Psi_p = \big\{(t,e) \,|\, 0 \leqslant t \leqslant \min\{l, |\mathbb{D}|\},\ 0 \leqslant e \leqslant \min\{t, m_p\}\big\}; \tag{6}$$

$$N_p(t,e) = C_{L-|\mathbb{D}|}^{l-t}\, H_{L-|\mathbb{D}|}^{l-t,m}(s_p(e)); \tag{7}$$

$$s_p(e) = \frac{l}{L}\big(n(a_p, \mathbb{X}) - \varepsilon(L-l)\big) - e.$$

Proof. We represent the probability of overfitting as

$$Q_\varepsilon = \sum_{p=0}^{P} \mathsf{P}\big[\mu X = a_p \text{ and } \delta(a_p, X) \geqslant \varepsilon\big].$$

We consider the set of the partitions $(X, \bar{X})$ for fixed values of t and e:

$$t = |X \cap \mathbb{D}|, \quad e = n(a_p, X \cap \mathbb{D}). \tag{8}$$

According to (6), the set of admissible values (t, e) is Ψ_p.

For such partitions, the validity of the condition $\delta(a_p, X) \geqslant \varepsilon$ is independent of the choice of the partition of the set $\mathbb{D}$, while by Lemma 1, the validity of the condition $\mu X = a_p$ is independent of the choice of the partition of the set $\mathbb{N}$ since the classifiers are independent of the set $\mathbb{N}$. This is why for each triple of the parameters p, t, e, the number of the partitions of the set $\mathbb{X}$ such that the conditions $\mu X = a_p$ and $\delta(\mu X, X) \geqslant \varepsilon$ hold true simultaneously is equal to the product $N_p(t, e)D_p(t, e)$.

Let us prove (7). Let $n(a_p, X \cap \mathbb{N}) = s$, then $n(a_p, X) = e + s$. The condition $\delta(a_p, X) \geqslant \varepsilon$ is equivalent to the condition $n(a_p, X) \leqslant \frac{l}{L}(n(a_p, \mathbb{X}) - \varepsilon(L - l))$, and thus, $s \leqslant s_p(e)$. Given t and s, the number of the partitions of the set $\mathbb{N}$ is equal to $C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s}$, which implies

$$N_p(t, e) = \sum_{s=0}^{s_p(e)} C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s} = C_{L-|\mathbb{D}|}^{l-t} \frac{1}{C_{L-|\mathbb{D}|}^{l-t}} \sum_{s=0}^{s_p(e)} C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s} = C_{L-|\mathbb{D}|}^{l-t} H_{L-|\mathbb{D}|}^{l-t,m}(s_p(e)).$$

□

For the functional of the complete cross-validation, a similar theorem holds.

Theorem 3. *For an arbitrary family of the classifiers $\mathbb{A} = \{a_0, \ldots, a_P\}$, a finite learning algorithm μ, a set $\mathbb{X}$ of a cardinality L, a size l of training sample, the functional of the complete cross-validation is of the form*

$$CCV = \frac{1}{(L - l)C_L^l} \sum_{p=0}^{P} \sum_{(t,e) \in \Psi_p} D_p(t, e) F_p(t, e), \tag{9}$$

where

$$F_p(t, e) = \sum_{s=0}^{\min\{l-t, m\}} C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s} \big(n(a_p, \mathbb{X}) - s - e\big), \tag{10}$$

the sets $\mathbb{D}$ and Ψ_p are determined by (3) *and* (6)*, the parameters m_p and m are determined by* (4)*.*

Proof. We write the formula for the complete cross-validation and interchange the summations signs:

$$CCV = \frac{1}{C_L^l} \sum_{X \in [X]^l} \sum_{p=0}^{P} [\mu X = a_p] \, \nu(a_p, \bar{X}) = \frac{1}{C_L^l} \sum_{p=0}^{P} \sum_{X \in [X]^l} [\mu X = a_p] \, \nu(a_p, \bar{X}).$$

By Lemma 1, the validity of the condition $\mu X = a_p$ is independent of the choice of the partition of the set $\mathbb{N}$.

We represent the number of errors of the classifier a_p on a testing sample as

$$n(a_p, \bar{X}) = n(a_p, \mathbb{X}) - n(a_p, X) = n(a_p, \mathbb{X}) - n(a_p, X \cap \mathbb{D}) - n(a_p, X \cap \mathbb{N}).$$

We define the parameters t and e by formulae (8). We denote $s = n(a_p, X \cap \mathbb{N})$. The restrictions $s + t \leqslant l$ and $s \leqslant m$ imply the upper bound for the parameter s in (10).

It is easy to check that the number of the partitions of the set $\mathbb{N}$ for given t and s is equal to $C_m^s C_{L-|\mathbb{D}|-m}^{l-t-s}$. This completes the proof. □

Thus, the problem is reduced to calculating $D_p(t, e)$ for each p on the entire set Ψ_p. In what follows, in the case of a direct sequence, we describe a recurrent algorithm for calculating $D_p(t, e)$.

3. Calculation of the number of partitions of the edges in direct sequence

Assume that the family $\mathbb{A} = \{a_0, \ldots, a_P\}$ is a direct sequence. We call the objects of the set $\mathbb{D}$ *edges of the direct sequence* $\mathbb{A}$.

3.1. Reduction to the problem on left and right sequences. We consider a classifier a_p and fix a point $(t, e) \in \Psi_p$. With respect to a_p, the direct sequence $\mathbb{A}$ is partitioned into two sequences, the left one $a_0, a_1, \ldots, a_p$ and the right one $a_p, a_{p+1}, \ldots, a_P$.

We are going to reduce the issue on calculating $D_p(t, e)$ to finding the number of partitions with some restrictions of the set of edges for the left and the right sequence.

Theorem 4. *Let μ be a finite learning algorithm. For each p, for all $(t, e) \in \Psi_p$ the number of partitions of the set $\mathbb{D}$ such that $t = |X \cap \mathbb{D}|$, $e = n(a_p, X \cap \mathbb{D})$ and $\mu X = a_p$ is equal to*

$$D_p(t, e) = \sum_{t'+t''=t} \sum_{e'+e''=e} L_p(t', e') R_p(t'', e''), \tag{11}$$

where

$$L_p(t', e') = \#\left\{ (X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p) \;\middle|\; \begin{array}{l} \forall d = 0, \ldots, p \;\; a_p \succ_X a_d, \\ t' = |X \cap \mathbb{L}_p|, \;\; e' = n(a_p, \, X \cap \mathbb{L}_p) \end{array} \right\}, \tag{12}$$

$$R_p(t'', e'') = \#\left\{ (X \cap \mathbb{R}_p, \bar{X} \cap \mathbb{R}_p) \;\middle|\; \begin{array}{l} \forall d = p+1, \ldots, P \;\; a_p \succ_X a_d, \\ t'' = |X \cap \mathbb{R}_p|, \;\; e'' = n(a_p, \, X \cap \mathbb{R}_p) \end{array} \right\}, \tag{13}$$

the sets $\mathbb{L}_p$ and $\mathbb{R}_p$ are the sets of the edges of the left and right sequences, respectively, the point (t', e') and (t'', e'') are the elements of the sets Ψ'_p and Ψ''_p, respectively, where

$$\Psi'_p = \big\{ (t', e') \;\big|\; 0 \leqslant t' \leqslant \min\{l, |\mathbb{L}_p|\}, 0 \leqslant e' \leqslant \min\{t', \, n(a_p, \mathbb{L}_p)\} \big\}, \tag{14}$$

$$\Psi''_p = \big\{ (t'', e'') \;\big|\; 0 \leqslant t'' \leqslant \min\{l, |\mathbb{R}_p|\}, 0 \leqslant e'' \leqslant \min\{t'', \, n(a_p, \mathbb{R}_p)\} \big\}. \tag{15}$$

Proof. The sets $\mathbb{L}_p$ and $\mathbb{R}_p$ do not intersect and hence, the classifiers of the left sequences are indiscernible on $\mathbb{R}_p$, the classifiers of the right sequence are indiscernible on $\mathbb{L}_p$. Then by Lemma 1, the validity of condition (2) is independent of the choice of the partition of the set $\mathbb{R}_p$. In the same way, the validity of condition (2) for all classifiers of the right sequence in independent of the choice of the partition of the set $\mathbb{L}_p$. Thus, the total number of partitions of the set $\mathbb{D}$, in which the learning algorithm chooses a_p, is the product of the number of the partitions of the sets $\mathbb{L}_p$ and $\mathbb{R}_p$, in which a_p is better than all classifiers of the left and right sequence, respectively. The parameters t', t'', e', e'' are necessary to satisfy the conditions defined by the parameters t and e. □

The partitions of the sets $\mathbb{L}_p$ and $\mathbb{R}_p$ satisfying conditions (12) and (13), respectively, will be called *admissible*.

We consider the algorithm of PERM. Let us prove that it is finite; then Theorems 2–4 hold for this method and for each p the problem is reduced to calculating the number of admissible partitions $L_p(t', e')$ and $R_p(t'', e'')$ for all points in the sets Ψ'_p and Ψ''_p.

We assume that among the classifiers minimizing the number of the errors on a training sample X and making the same number of errors on a testing sample $\bar{X}$, we choose the classifier with the maximal index. This restriction makes no influence on the estimate for the probability of overfitting and of complete cross-validation but allows us to calculate precisely the required number of the partitions.

Definition 5. *The error reserve of a classifier a with respect to a_p on a sample X is the quantity $\Delta_p(a, X) = n(a, X) - n(a_p, X)$.*

Lemma 2. *The algorithm of PERM is finite with the order relation $\succ_X$ defined as follows: a classifier a_p is better than a classifier a on a sampling X if and only if one the following conditions holds:*

1) $\Delta_p(a, X) > 0$;
2) $\Delta_p(a, X) = 0$ *and* a *is contained in the left sequence and* $n(a, \mathbb{X}) \leqslant n(a_p, \mathbb{X})$;
3) $\Delta_p(a, X) = 0$ *and* a *is contained in the right sequence and* $n(a, \mathbb{X}) < n(a_p, \mathbb{X})$.

This lemma is implied by the definition of PERM.

In what follows we consider the case, when the direct sequence $\mathbb{A}$ is a direct chain. Then the left and right sequences $\mathbb{L}_p$ and $\mathbb{R}_p$ are also direct chains. We consider the algorithm of PERM μ with the order relation $\succ_X$ defined by Lemma 2.

3.2. Finding the number of admissible partitions of the edges in the left chain. Let us find $L_p(t', e')$ for each p in each point $(t', e') \in \Psi'_p$. We observe that as $p = 0$, this problem can be solved trivially: the set Ψ'_0 consists of the single point $(0, 0)$ and $L_0(0, 0) = 1$. Hereafter we assume $1 \leqslant p \leqslant P$.

We renumber the classifiers so that the sequence begins at a_p and ends at a_0. We denote $\{b_0, \ldots, b_p\}$, where $b_d = a_{p-d}$ for each $d = 0, \ldots, p$. We write the reserve of the error with respect to a_p as $\Delta_0(b_d, X) = \Delta_p(a_{p-d}, X)$ for each d.

The left chain $\mathbb{L}_p$ is formed by increasing and decreasing monotone segments. We denote the set of the edges of increasing segments by $\mathbb{C}_p$, while $\mathbb{I}_p$ stands for the monotone segments of the chain. We have $\mathbb{C}_p \sqcup \mathbb{I}_p = \mathbb{L}_p$.

The chain is direct and therefore, b_0 makes no error on all objects $\mathbb{C}_p$, that is,

$$\mathbb{C}_p = \{x \in \mathbb{L}_p \colon I(b_0, x) = 0\}, \quad \mathbb{I}_p = \{x \in \mathbb{L}_p \colon I(b_0, x) = 1\}. \tag{16}$$

Then the identity $e' = |X \cap \mathbb{I}_p|$ holds and $|X \cap \mathbb{C}_p| = t' - e'$.

We note that since the classifiers of the left chain are discernible only on the objects of the set $\mathbb{L}_p$, for each classifier b in the left chain we have

$$\Delta_0(b, X) = \Delta_0(b, X \cap \mathbb{L}_p), \quad \forall X \subseteq \mathbb{X}.$$

This implies that fixing a partition of the set $\mathbb{L}_p$, we determine the reserve of the errors on all corresponding training samples X.

We introduce the three-dimensional lattice

$$\Omega_p = \{0, \ldots, |\mathbb{L}_p|\} \times \{-|\mathbb{L}_p|, \ldots, |\mathbb{L}_p|\} \times \{0, \ldots, |\mathbb{L}_p|\}.$$

Definition 6. *On* Ω_p *we define the set* $\mathbb{T}_p$ *of the paths leaving the point* $(0, 0, 0)$ *and formed by the steps of three types:*

1) "right", from the point (d, Δ, i) *to the point* $(d + 1, \Delta, i)$;
2) "right and up", from the point (d, Δ, i) *to the point* $(d + 1, \Delta + 1, i)$;
3) "right and down", from the point (d, Δ, i) *to the point* $(d + 1, \Delta - 1, i + 1)$;

and for each d *the step from the point* (d, Δ, i) *satisfies the condition: assume that the classifiers* b_d *and* b_{d+1} *are connected by an edge* x*, then*

1) if $x \in \mathbb{C}_p$*, this is a step "right" or "right and up";*
2) if $x \in \mathbb{I}_p$*, this is a step "right" or "right and down".*

Theorem 5. *There exists a one-to-one correspondence between the partitions of the set* $\mathbb{L}_p$ *and the paths in the set* $\mathbb{T}_p$*. The paths corresponding to the partition* $(X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p)$ *passes the points* (d, Δ, i)*, where for each* $d = 0, \ldots, p$ *we have* $\Delta = \Delta_0(b_d, X)$*, and the coordinate* i *is equal to the number of the edges in* $X \cap \mathbb{I}_p$ *between* b_0 *and* b_d*.*

Proof. Suppose that the classifiers b_{d-1} and b_d are connected by an edge x. If $x \in \bar{X}$, then $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X)$ since the reserve of the errors depends on X only.

Assume that x is an element of X. If x is contained in an increasing chain, then b_{d-1} makes no error on this edge, while b_d does. Then $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) + 1$. If x is contained in $\mathbb{I}_p$, then b_{d-1} makes an error on this object, while b_d does not. Hence, $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) - 1$.

To the partition of the set $\mathbb{L}_p$, we associate a path by the following rule. Let a path pass the point (d, Δ, i). As $d = 0$, we assume that this is the point $(0, 0, 0)$. From this point along this path, the step is of the form "right" if $x \in \bar{X}$; "right and up" if $x \in X \cap \mathbb{C}_p$; "right and down" if $x \in X \cap \mathbb{I}_p$.

Then for each d the coordinates Δ and i make the sense described in the formulation of the theorem and in the described steps they change at most by 1. Hence, the path is located in the lattice Ω_p and therefore, in the set $\mathbb{T}_p$ and it is determined uniquely.

By the same rule, to each path in $\mathbb{T}_p$, we associate uniquely a partition of the set $\mathbb{L}_p$. Hence, the mapping from the set of partitions into the set of the paths is surjective and injective and hence, is bijective. □

Example 4. *In Figure 2, the lower graph demonstrates a chain and its edges contained in the training sample are highlighted by the double line. To such partition of the edges of the chain, a path is associated and its projection on the plane (d, Δ) is shown in the upper graph. In this example, the path passes the points with a negative coordinate Δ. Hence, the chain contains the classifiers with a negative reserve of the errors. Therefore, by Lemma 2 and by conditions* (2)*, under such partition the classifier b_0 is not chosen by the learning algorithm. Excluding the paths not obeying Lemma 2 from the consideration, we exclude also non-admissible partitions.*

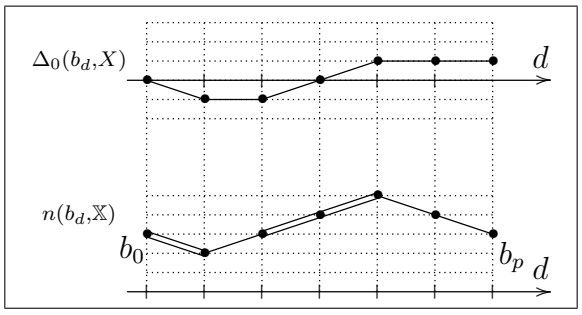


FIGURE 2. Correspondence of the partition of a chain (lower graph) to the projection of the path (upper graph). By the double line, we highlight the edges in the training sample.

We introduce the set

$$\Omega'_p = \left\{ (d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ and } |\Delta| \leqslant d \text{ and} \\ \big(\text{or } \Delta > 0, \text{ or } (\Delta = 0 \text{ and } n(b_d, \mathbb{X}) \leqslant n(b_0, \mathbb{X}))\big) \end{array} \right\}. \tag{17}$$

Lemma 3. *Each point (d, Δ, i) of the path in $\mathbb{T}_p$ corresponding to an admissible partition of the set $\mathbb{L}_p$ belongs to the set $\Omega'_p \subseteq \Omega_p$.*

Proof. First two conditions in Definition (17) are due to Theorem 5. The third condition repeats the condition of Lemma 2. □

Let $T_p(d, \Delta, i)$ be the number of the paths in $\mathbb{T}_p$ connecting the point $(0, 0, 0)$ with (d, Δ, i) passing only the points in the set Ω'_p. The rules of constructing the path by a partition of the set $\mathbb{L}_p$ imply the following lemma.

Lemma 4. *At each point (d, Δ, i) in the three-dimensional lattice Ω_p, the quantity $T_p(d, \Delta, i)$ is calculated recurrently:*

1) The initial condition is $T_p(0, 0, 0) = 1$.
2) If $(d, \Delta, i) \notin \Omega'_p$, then $T_p(d, \Delta, i) = 0$.

3) Assume that b_{d-1} and b_d are connected by the edge x. Then

$$T_p(d,\,\Delta,\,i) = \begin{cases} T_p(d-1,\,\Delta,\,i) + T_p(d-1,\,\Delta-1,\,i) & \text{if} \quad x \in \mathbb{C}_p, \\ T_p(d-1,\,\Delta,\,i) + T_p(d-1,\,\Delta+1,\,i-1) & \text{if} \quad x \in \mathbb{I}_p, \end{cases} \tag{18}$$

where the sets $\mathbb{C}_p$ and $\mathbb{I}_p$ are determined by (16).

Theorem 6. *Assume that we are given an algorithm of PERM μ, a set $\mathbb{X}$ of a cardinality L, a size l of a training sample and a direct chain $\mathbb{A} = \{a_0, \ldots, a_P\}$. Then for each $p = 1, \ldots, P$ at each point (t', e') of the set Ψ'_p defined in* (14)*, the number $L_p(t', e')$ of admissible partitions of the set $\mathbb{L}_p$ determined by* (12) *is equal to*

$$L_p(t',\,e') = T_p(|\mathbb{L}_p|,\,t'-2e',\,e')$$

and is calculated recurrently by the rules described in Lemma 4, where $b_d = a_{p-d}$ for each d, under the boundary conditions $L_0(0,0) = 1$.

Proof. It follows from Theorem 5 that

$$\Delta_p(a_0, X) = |X \cap \mathbb{C}_p| - |X \cap \mathbb{I}_p| = t' - 2e'.$$

There exists a bijection between the set of the edges of the left chain and the paths in $\mathbb{T}_p$. Thus, the number of the paths passing point $(p,\,t'-2e',\,e')$ is equal to the number of the partitions satisfying the conditions $t' = |X \cap \mathbb{L}_p|$ and $e' = n(a_p, X \cap \mathbb{L}_p)$. Keeping only the paths passing the points in the set $\Omega'_p(t', e')$, we keep only those associated with the admissible partitions. Their number is equal to $T_p(|\mathbb{L}_p|,\,t'-2e',\,e')$. □

Remark 1. *The restrictions $i \leqslant e'$ and $\Delta \leqslant t' - e'$ implied by Theorem 5 hold immediately for the paths connecting the points $(0,0,0)$ and $(p, t'-2e', e')$. Indeed, since the values i and $\Delta + i$ do not increase, they do not exceed the values at the end point, that is, $i \leqslant e'$ and*

$$\Delta + i \leqslant t' - 2e' + e' = t' - e'.$$

We have $i \geqslant 0$ and hence, $\Delta \leqslant \Delta + i \leqslant t' - e'$. In view of this fact, the definition of the set Ω'_p do not involve these restrictions.

Thus, we have learned how to solve the problem for the left chain.

3.3. Finding admissible partitions for the set of the edges of the right chain. We want to calculate $R_p(t'', e'')$ for each p at each point $(t'', e'') \in \Psi''_p$. The procedure reproduces almost literally that for the left chain after changing $\mathbb{L}_p$ by $\mathbb{R}_p$ and the point (t', e') by (t'', e''). We also have the boundary conditions: as $p = P$, we have $\Psi''_P = \{(0,0)\}$ and $R_P(0,0) = 1$. We let $0 \leqslant p \leqslant P-1$.

For each $d = 0, \ldots, P-p$, by $b_d = a_{p+d}$ we denote the classifiers in the chain. It follows from Lemma 2 that Lemma 4 is true for the right chain once we replace the set Ω'_p by the set Ω''_p defined as

$$\Omega''_p = \left\{ (d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ and } |\Delta| \leqslant d \text{ and} \\ \big(\text{either } \Delta > 0, \text{ or } (\Delta = 0 \text{ and } n(b_d, \mathbb{X}) < n(b_0, \mathbb{X}))\big) \end{array} \right\}. \tag{19}$$

Similar to Theorem 6, we have the following theorem for the right chain.

Theorem 7. *Assume that we are given an algorithm of PERM μ, a set $\mathbb{X}$ of a cardinality L, a size l of training sample l and an arbitrary direct chain $\mathbb{A} = \{a_0, \ldots, a_P\}$. Then for each $p = 0, \ldots, P-1$ at each point (t'', e'') of the set Ψ''_p introduced in* (15)*, the number $R_p(t'', e'')$ of admissible partitions of the set $\mathbb{R}_p$ defined by (13) is equal to*

$$R_p(t'', e'') = T_p(|\mathbb{R}_p|, t'' - 2e'', e'')$$

and it is calculated recurrently by the rules described in Lemma 4 with the set Ω'_p replaced by Ω''_p and b_d replaced by a_{p+d} for each d. The boundary value are $R_P(0,0) = 1$.

Remark 2. *By Lemma 2, for each $d = 0, \ldots, P$ the reserve of errors for the classifier a_d of the chain can be negative for admissible partitions of the set of the edges of the left and right chain. In particular, $\Delta_p(a_0, X) = t' - 2e' \geqslant 0$ and $\Delta_p(a_P, X) = t'' - 2e'' \geqslant 0$. Hence, the second coordinates of the points in the sets Ψ_p, Ψ'_p, Ψ''_p range as*

$$0 \leqslant e \leqslant \min\bigl\{\tfrac{1}{2}t,\, m_p\bigr\}, \qquad 0 \leqslant e' \leqslant \min\bigl\{\tfrac{1}{2}t', n(a_p, \mathbb{L}_p)\bigr\}, \qquad 0 \leqslant e'' \leqslant \min\bigl\{\tfrac{1}{2}t'', n(a_p, \mathbb{R}_p)\bigr\}.$$

3.4. Finding the number of admissible partitions of the set of edges in direct sequence. We consider the general case of a direct sequence $\mathbb{A} = \{a_0, \ldots, a_P\}$. We are going to reduce the problem on calculating the number of admissible partitions for the left and right sequences to similar problems for direct chains.

In order to do this, we construct a direct chain $\mathbb{A}_c$ such that $\mathbb{A} \subseteq \mathbb{A}_c$ and the right and the last classifiers in the family coincide. We do this as follows: for each i such that $|G_i| > 1$, we add the direct chain $\mathbb{G}_i$ into the sequence $\mathbb{A}$

$$\{a_0, \ldots, a_{i-1}\} \cup \mathbb{G}_i \cup \{a_{i+2}, \ldots, a_P\},$$

where the direct chain $\mathbb{G}_i$ is such that the first classifier of the chain is a_i and the last classifier is a_{i+1}. For the sake of definiteness we assume that $\mathbb{G}_i$ is constructed as a direct chain formed by two monotone ones: a decreasing chain of length n_1 and an increasing chain of length n_0, where

$$n_1 = \#\{x \in G_i \,|\, I(a_i, x) = 1\}, \quad n_0 = \#\{x \in G_i \,|\, I(a_i, x) = 0\}.$$

We call the constructed chain $\mathbb{A}_c$ *interpolation* of the chain $\mathbb{A}$. Its length is equal to $|\mathbb{D}|$.

For each $a_p \in \mathbb{A}$ we consider the left sequence $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}$ and the left chain $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}_c$. By construction, the sets of the edges in these families coincide and hence, the sets of the admissible partitions of the left chain and of the left sequence defined by (12) also coincide. Let us calculate their number by Theorems 6 and 7, up to a single difference.

According to (2), the condition $a_p \succ_X a$ should hold only for $a \in \mathbb{A}$. This restriction determines the structure of the sets Ω'_p and Ω''_p defined in (17) and (19). We redefine them for the interpolation of the sequence $\mathbb{A}$:

$$\Omega'_p = \left\{(d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ or } \bigl(b_d \in \mathbb{A} \text{ and } 0 \leqslant i \leqslant d \text{ and } |\Delta| \leqslant d \\ \qquad \text{and } \bigl(\Delta > 0 \text{ or } (\Delta = 0 \text{ and } n(b_d, \mathbb{X}) \leqslant n(b_0, \mathbb{X}))\bigr)\bigr) \end{array}\right\}; \tag{20}$$

$$\Omega''_p = \left\{(d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ or } \bigl(b_d \in \mathbb{A} \text{ and } 0 \leqslant i \leqslant d \text{ and } |\Delta| \leqslant d \\ \qquad \text{and } \bigl(\Delta > 0 \text{ or } (\Delta = 0 \text{ and } n(b_d, \mathbb{X}) < n(b_0, \mathbb{X}))\bigr)\bigr) \end{array}\right\}. \tag{21}$$

Theorem 8. *Assume that we are given an algorithm of PERM μ, a set $\mathbb{X}$ of a cardinality L, a size l of a training sample and a direct sequence $\mathbb{A} = \{a_0, \ldots, a_P\}$. Let the direct chain $\mathbb{A}_c = \{c_0, \ldots, c_{|\mathbb{D}|}\}$ be an interpolation of the sequence $\mathbb{A}$. To each classifier $a_p \in \mathbb{A}$, there corresponds a classifier $c_{i_p} \in \mathbb{A}_c$.*

Then for each $p = 1, \ldots, P$ at each point (t', e') of the set Ψ'_p defined in (14), the number $L_p(t', e')$ of admissible partitions of the set $\mathbb{L}_p$ defined by (12) is equal to

$$L_p(t',\, e') = T_p(|\mathbb{L}_p|,\, t' - 2e',\, e') \tag{22}$$

and is calculated recurrently by the rules described in Lemma 4, where $b_d = c_{i_p - d}$ for each d and the set Ω'_p is defined by (20). The boundary conditions are $L_0(0, 0) = 1$.

For each $p = 0, \ldots, P-1$ at each point (t'', e'') of the set Ψ''_p defined in (15), the number $R_p(t'', e'')$ of admissible partitions of the set $\mathbb{R}_p$ defined by (13) is equal to

$$R_p(t'', e'') = T_p(|\mathbb{R}_p|, t'' - 2e'', e'') \tag{23}$$

and is calculated recurrently by the rules described in Lemma 4 with replacing the set Ω'_p by Ω''_p defined in (21) and with replacing b_d by $c_{i_p + d}$ for each d. The boundary conditions are $R_P(0, 0) = 1$.

4. Algorithm of calculating the probability of overfitting and the complete cross-validation

Thus, Theorem 8 describes an algorithm for finding the number of admissible partitions of the sets of the edges in the left and the right sequence for each p. It remains to substitute the found values into formulae (11), (5) and (9). To reduce the calculations by Theorems 2 and 3, for each p, it is proposed to calculate in advance $L_p(t', e')$, $R_p(t'', e'')$, $N_p(t, e)$ and $F_p(t, e)$ and to sum up the obtained values. The scheme of calculations is shown in algorithm 1.

Algorithm 1: Calculation of the probability of overfitting and of complete cross-validation

Input: error matrix of the direct sequence $\mathbb{A} = \{a_0, \ldots, a_P\}$, parameters l, ε.
Output: probability of overfitting Q_ε and complete cross-validation CCV.

1 construct the direct chain $\mathbb{A}_c$ being an interpolation of the sequence $\mathbb{A}$;
2 find m by (4);
3 **for all** $p = 0, \ldots, P$
4 partition the chain $\mathbb{A}_c$ into two chains, the left one $\{a_p, \ldots, a_0\}$ and the right one $\{a_p, \ldots, a_P\}$;
5 **for all** *points* (t', e') *in the set* Ψ'_p *defined by* (14)
6 find $L_p(t', e')$ by formulae (22), (18) and (20);
7 **for all** *points* (t'', e'') *in the set* Ψ''_p *defined by* (15)
8 find $R_p(t'', e'')$ by formulae (23), (18) and (21);
9 **for all** *points* (t, e) *in the set* Ψ_p *defined by* (6)
10 calculate $N_p(t, e)$ by formula (7);
11 calculate $F_p(t, e)$ by formula (10);
12 $Q_\varepsilon := \dfrac{1}{C_L^l} \sum\limits_{p=0}^{P} \sum\limits_{(t',e')\in\Psi'_p} \sum\limits_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') N_p(t' + t'', e' + e'')$;
13 $CCV := \dfrac{1}{(L-l)C_L^l} \sum\limits_{p=0}^{P} \sum\limits_{(t',e')\in\Psi'_p} \sum\limits_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') F_p(t' + t'', e' + e'')$;

4.1. Complexity of the algorithm. Let us estimate the complexity of steps 5–11 of algorithm 1.

Calculating $L_p(t', e')$ by Theorem 6 at steps 5–6, we calculate $T_p(d, \Delta, i)$ once for all $(d, \Delta, i) \in \Omega'_p$, then for each $(t', e') \in \Psi'_p$, we let the quantity $L_p(t', e')$ to be equal to $T_p(d, t' - 2e', e')$. The set Ω'_p is embedded into the cube with a side of the length $O(|\mathbb{L}_p|)$ since the absolute value of each coordinate is bounded by the number of the edge in the left sequence. Therefore, the complexity of steps 5–6 is $O(|\mathbb{L}_p|^3)$. In the same way, the complexity of steps 7–8 is $O(|\mathbb{R}_p|)^3)$.

In order to find $N_p(t, e)$ and $F_p(t, e)$, we need to calculate the binomial coefficients C_m^i and C_{L-P-m}^i for all possible i in $O(L)$. The binomial coefficients are not recalculated for all p. Under the known values of the binomial coefficients, the sought values $N_p(t, e)$ and $F_p(t, e)$ are calculated in $O(L)$. The set Ψ_p is embedded into the square with the side of the length L, and hence, the steps 9–11 are made in $O(L^3)$. Therefore, the complexity of steps 5–11 is $O(|\mathbb{D}|^3 + L^3) = O(L^3)$ for each p.

The sets Ψ'_p and Ψ''_p are embedded into the square with the side of the length P and hence, steps 12–13 are made in $O(L^5)$ and the complexity of algorithm 1 is also $O(L^5)$.

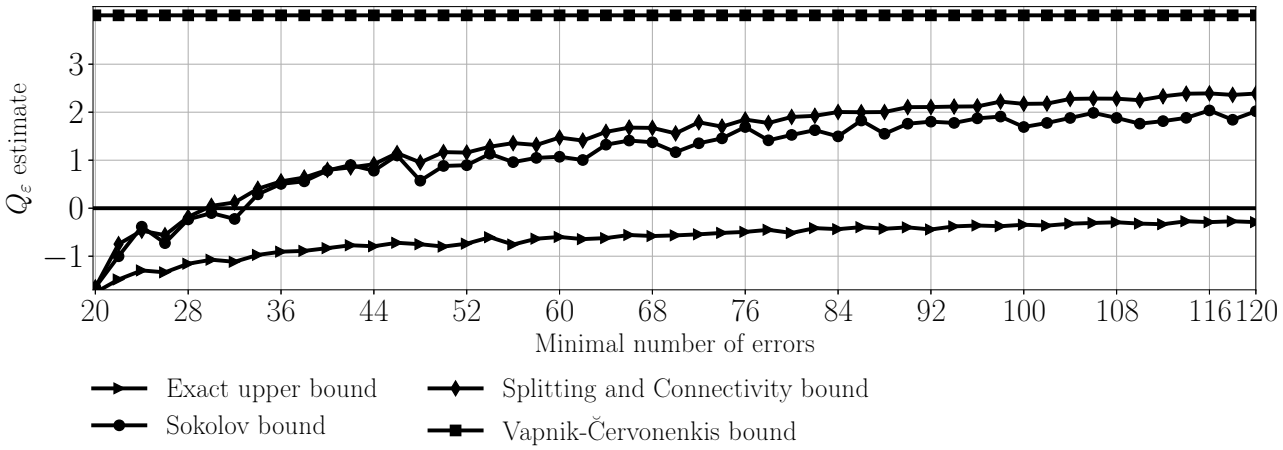


FIGURE 3. Comparison of upper bounds for the probability of overfitting in the logarithmic scale. The horizontal line stands for the value $Q_\varepsilon = 1$. The experiment conditions are $L = 240$, $\ell = 160$, $m = 20$, $\varepsilon = 0.05$. The horizontal direction indicates the minimal number of the errors of the classifiers.

5. Comparison with the known estimates for the probability of overfitting

Let us consider the family of one-dimensional threshold rules in the classification problem for the classes of equal cardinality. Let us show that for this problem the known upper bounds for the probability of overfitting are overstated.

In figure 3, in the logarithmic scale, there shown the Vapnik–Červonenkis bounds [1], the splitting and connectivity bounds [12] and Sokolov bounds [17] in comparison with the sharp upper bound for the probability of overfitting of a direct sequence. The splitting and connectivity bounds and Sokolov bound are sharp in the only case, when the minimal number of errors coincide with the parameter m. In this case the boundary between the classes is found exactly and the family is a unimodal chain [9]. As the minimal number of errors grows, the Sokolov bound exceeds the sharp upper bound. The Vapnik–Červonenkis estimates for the considered sequence turn out to be overstated for any value of the minimal number of the errors.

6. Conclusion

We introduce the notion of the a finite learning algorithm, for which we develop the algorithm for calculating the probability of overfitting and of complete cross-validation of direct sequences of the classifiers generated by elementary threshold rules as the threshold parameter varies. We show that the algorithm of empirical risk minimization (ERM) and the algorithm of overfitting maximization (OM) are finite. For ERM we show that the known upper bounds for the probability of overfitting of direct sequences are overstated and are not applicable for practical problems.

A future issue for studying is application of this algorithm for increasing the learning ability of the methods of statistical learning, in particular, for improving criteria of choosing features, for the methods of seeking logical laws in data, for linear and logical classification algorithms. Another direction is a generalization of this algorithm for other functional of generalizing ability, in particular, for the functional of expected overfitting of an algorithm of OM, which is equal to the Rademacher complexity of a family and which connects the combinatorial theory of overfitting with the theory of empirical processes and the theory of inequalities for a concentration of a probability measure.

The author is deeply grateful to his scientific supervisor K.V. Vorontsov for a permanent attention to the work and valuable remarks.

BIBLIOGRAPHY

1. V.N. Vapnik, A.Ya. Červonenkis. *On uniform convergence of the frequencies of events to their probabilities* // Teor. Veroyat. Primen. **16**:2, 264–280 (1971). [Theor. Prob. Appl. **16**:2, 264–280 (1971).]
2. S. Boucheron, O. Bousquet, G. Lugosi. *Theory of classification: A survey of some recent advances* // ESAIM: Prob. Stat. **9**:, 323–375 (2005).
3. V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems.* École d'été de probabilités de Saint-Flour XXXVIII-2008. Lecture Notes Math. Springer, Heidelberg (2011).
4. K. V. Vorontsov. *Combinatorial probability and the tightness of generalization bounds* // Pattern Recogn. Image Anal. **18**:2, 243–259 (2008).
5. D. Haussler, N. Littlestone, M.K. Warmuth. *Predicting* $\{0,1\}$*-functions on randomly drawn points* // Inf. Comput. **115**:, 248–292 (1994).
6. K.V. Vorontsov. *Combinatorial bounds for learning performance* // Dokl. RAN. **394**:2, 175–178 (2004). [Dokl. Math. **69**:1, 145–148 (2004).]
7. K.V. Vorontsov. *Tight bounds for the probability of overfitting* // Dokl. RAN. **429**:1, 15–18 (2009). [Dokl. Math. **80**:3, 793–796 (2009).]
8. K.V. Vorontsov. *Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting* // Pattern Recogn. Image Anal. **19**:3, 412–420 (2009).
9. K.V. Vorontsov. *Exact combinatorial bounds on the probability of overfitting for empirical risk minimization* // Pattern Recogn. Image Anal. **20**:3, 269–285 (2010).
10. V. Koltchinskii. *Rademacher penalties and structural risk minimization* // IEEE Trans. Inf. Theory. **47**:5, 1902–1914 (2001).
11. K.V. Vorontsov. *Combinatorial theory of overfitting: how connectivity and splitting reduces the local complexity* // Presentation at "9th IFIP WG 12.5 International conference", Paphos, Cyprus (2013).
12. K.V. Vorontsov, A.A. Ivahnenko. *Tight combinatorial generalization bounds for threshold conjunction rules* // in Proc. "4th International conference on Pattern Recognition and Machine Intelligence". Lecture Notes Comp. Sci. Springer, Berlin, 66–73 (2011).
13. N.K. Zhivotovskii, K.V. Vorontsov. *Sharpness criterions in combinatorial estimates for generalizing ability* // in Proc. "Intellectualization of data processing", Torus Press, Moscow, 25–28 (2012). (in Russian).
14. Yu.I. Zhuravlev, V.V. Ryazanov, O.V. Senko. *"Recognition". Mathematical methods. Program environment. Practical applications.* Fazis, Moscow (2006). (in Russian).
15. Yu.I. Zhuravlev. *On algebraic approach to solving recognition or classification problems* // Probl. Kibern. **33**, 5–68 (1978). (in Russian).
16. I.S. Guz. *Constructive evaluation of the complete cross-validation for threshold classification* // Matem. Biol. Bioinform. **6**:2, 173–189 (2011). (in Russian).
17. K.V. Vorontsov, A.I. Frey, E.A. Sokolov. *Calculable combinatorial estimates for probability of overfitting* // Mashin. Obuch. Anal. Dannych. **1**:6, 734–743 (2013). (in Russian).
18. A.I. Frey, I.O. Tolstikhin. *Combinatorial estimates of probability of overfitting on the base of clustering and covering the set of algorithms* // Mashin. Obuch. Anal. Dannych. **1**:6, 761–778 (2013). (in Russian).
19. A.I. Frey, I.O. Tolstikhin. *Cover-based combinatorial bounds on probability of overfitting* // Dokl. RAN. **455**:3, 265–268 (2014). [Dokl. Math. **89**:2, 185–187 (2014).]

Shaura Khabirovna Ishkina,
Federal Research Center "Computer Science and Control" of RAS,
Vavilova str. 44/2
119333, Moscow, Russia
E-mail: `shaura-ishkina@yandex.ru`