Splitting and Similarity Phenomena in the Sets of Classifiers and Their Effect on the Probability of Overfitting

K. V. Vorontsov

Dorodnitsyn Computing Centre, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia
e-mail: voron@ccas.ru

Abstract—It is shown that computationally tight bounds for the probability of overfitting can be obtained only by simultaneous consideration of the following two properties of classifier sets: splitting into error levels and similarity of classifiers. For a set consisting of only two classifiers, an exact bound is obtained for the probability of overfitting. This is the simplest learning task that exhibits overfitting and the effects of splitting and similarity, which reduce the probability of overfitting. For a more complex case—a chain of classifiers—an experiment is carried out in which the effects of splitting and similarity are estimated separately. It is shown that reasonably low probabilities of overfitting can be obtained only for the sets that possess both properties.

DOI: 10.1134/S1054661809030055

One of the main problems in statistical learning theory is obtaining sufficiently tight generalization bounds. This problem has remained open for more than 40 years since the rise of Vapnik–Chervonenkis (VC) theory [15, 16]. The tightest of the known bounds are still highly overestimated [10]. The overestimation leads either to an unjustified requirement to increase the sample length up to 10^5–10^8 objects [17] or, in structural risk minimization, to oversimplification of classifiers [8]. The known bounds only qualitatively describe the relation between overfitting and the complexity of a set of classifiers; however, they do not always admit exact quantitative predictions and controlling the learning process. The question of whether or not overfitting is related to some finer and not-yet-studied phenomena remains open.

The aim of the present study is to find the cause of overestimation and to search for ways to improve the bounds. We show that the overfitting probability essentially depends not only on the complexity of the set (the number of classifiers in the set) but also on the diversity of these classifiers. To obtain tight bounds, one should simultaneously take into account the following two facts: similarity of classifiers in the set and splitting of the set into error levels. The neglect of one of these factors frustrates all the efforts to take into account the second factor. This conclusion is also confirmed by the argument that known attempts to take into account one of these factors separately [2, 1 4, 10] have not radically improved the bound.

In Section 1, we introduce necessary concepts and definitions, including the weak (permutational) probabilistic axiom. Section 2 is of survey character; in this section we present some improvements of VC bounds due to taking into account the diversity of classifiers. In Section 3, we derive an exact combinatorial bound for the probability of overfitting for a set of two classifiers. This is the simplest particular case that exhibits both overfitting and splitting and similarity properties, which reduce the probability of overfitting. In Section 4, we consider special classifier sets, called chains of classifiers, for which the effects of splitting and similarity can be estimated separately. Model experiments show that computationally tight bounds for the probability of overfitting can only be obtained by simultaneous consideration of the splitting and similarity of the classifiers.

1. PROBLEM OF ESTIMATING THE PROBABILITY OF OVERFITTING

Suppose given a set $\mathbb{X} = \{x_1, ..., x_L\}$, called a *full,* or *general*, sample. The elements of the set $\mathbb{X}$ are called *objects.* Let $\mathbb{A}$ be a set whose elements are called *classifiers.* There exists a binary loss function I: $\mathbb{A} \times \mathbb{X} \to \{0, 1\}$. If $I(a, x) = 1$, it is said that the classifier a makes an error on the object x.

The number of errors of a classifier a on a sample $X \subseteq \mathbb{X}$ is defined as

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

The *error rate,* or the *empirical risk*, of a classifier a on a sample X is the quantity $\nu(a, X) = \frac{1}{|X|} n(a, X)$; it takes values on the interval [0, 1].

Received February 20, 2009

Denote by $\mathbb{X}_L^l$ the set of all l-element subsets of the general sample $\mathbb{X}$. It is obvious that $|\mathbb{X}_L^l| = C_L^l$.

A *learning algorithm* is a function μ: $\mathbb{X}_L^l \to \mathbb{A}$ that maps a certain classifier $a = \mu X$ from $\mathbb{A}$ to an arbitrary *training sample* $X \in \mathbb{X}_L^l$.

Empirical risk minimization] is a learning algorithm

$$\mu X = \arg\min_{a \in \mathbb{A}} n(a, X). \qquad (1.1)$$

Let us explain these concepts. In classification problems, a classifier is a computable function a: $\mathbb{X} \to \mathbb{Y}$ that assigns a class label from a given finite set $\mathbb{Y}$ to each object x from $\mathbb{X}$; the error indicator is given by

$$I(a, x) = [a(x) \neq y(x)],$$

where y: $\mathbb{X} \to \mathbb{Y}$ is an unknown target function. Here and below, brackets are used to transform a logical variable into the numbers 0 or 1 according to Iverson's convention [true] = 1 and [false] = 0 [5]. The set $\mathbb{A}$ is a parametric set of classifiers, for example, separating hyperplanes, neural networks, decision trees, etc. [6]. A learning algorithm μ learns the parameters of the classifier from a given training sample X with known classifications $y_i = y(x_i)$. It is also said that the algorithm μ *learns the function* $y(x)$ from the empirical data $(x_i, y_i)_{i=1}^l$. Examples of well-known learning algorithms are support vector machines (SVMs) for separating hyperplanes, back propagation for neural networks, and C4.5 for decision trees [6].

In regression problems, "classifiers" are functions a: $\mathbb{X} \to \mathbb{R}$; the error indicator can be defined as

$$I(a, x) = [|a(x) - y(x)| \geq \delta],$$

where y: $\mathbb{X} \to \mathbb{Y}$ is an unknown regression function and δ is the error threshold.

In the present study, there is no need to specify what a classifier is. It suffices to assume that classifiers are elements of an abstract set $\mathbb{A}$ under the additional assumption that there exists a binary loss function that gives 1 iff a classifier a makes an error on an object x. This interpretation of a classifier, on the one hand, extends the class of problems considered but, on the other hand, restricts this class to problems in which the value of the error is not essential.

The *deviation of error rates* of a classifier a on two samples X and $\bar{X} = \mathbb{X}\backslash X$ is the difference $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

The deviation of error rates of a classifier $a = \mu(X)$ is called the *overfitting* of the algorithm μ on the sample X:

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

We will say that an algorithm μ is *overfitted* on a sample X if $\delta_\mu(X) \geq \varepsilon$, where ε is the threshold parameter.

Note that usually the term overfitting is introduced informally and denotes a frequently encountered unwanted phenomenon when a classifier learned from a training sample works much more poorly on new testing data. Here we give this term a more rigorous formal meaning.

We will stick to the *weak probabilistic axiom* [18], which is based on a single probabilistic assumption. It is assumed that all C_L^l partitions of the general sample $\mathbb{X}$ into an *observed* training sample X of length l and a *hidden* testing sample $\bar{X}$ of length $k = L - l$ can be realized with equal probability. This assumption is in fact equivalent to the standard conjecture that the elements of the sample $\mathbb{X}$ are independent. However, it is not assumed that the probability measure exists on the whole space of objects; moreover, even the space itself is not introduced. Under the weak axiom, events are subsets of partitions of the sample $\mathbb{X}$. More precisely, for an arbitrary predicate β: $\mathbb{X}_L^l \to \{true, false\}$, the probability of event $\beta(X)$ is defined as the fraction of partitions for which $\beta(X)$ is true:

$$\mathrm{P}[\beta(X)] = \frac{1}{C_L^l} \sum_{X \in x_L^l} [\beta(X)].$$

Within the weak axiom, we will consider one of the main problems of statistical learning theory. Our goal is to obtain tight upper bounds for the probability of overfitting for a given algorithm μ:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathrm{P}[\delta_\mu(X) \geq \varepsilon]. \qquad (3.1)$$

The introduction of the weak axiom is motivated by the following arguments.

First, in data analysis problems, samples may only be finite, no matter if these samples are observed historical data or hidden future data. In some problems, the number of predictions k is so small that it is merely incorrect to introduce the error probability as the limit of the error rate as $k \to \infty$. The weak axiom allows one to obtain exact nonasymptotic results, valid for any finite l and k, by purely combinatorial methods. The concept of error probability in the weak axiom is not defined at all. The quality of classifiers is characterized by their error rate on finite samples. The value of overfitting is defined as the difference of error rates on two subsamples, rather than the difference between the error rate and the error probability. Note that this approach is not new in statistical learning theory. The first studies by Vapnik and Chervonenkis [16] were also based on estimations of the difference of error rates in two subsamples.

Second, one can easily estimate empirically the probabilities defined in terms of the fractions of sample partitions by replacing the average over all partitions by the average over a random subset of partitions (the Monte Carlo method). This method resembles cross-validation [4, 9] but differs from it in that here one estimates the empirical distribution of overfitting, δ_μ, rather than the empirical mean of the error rate on a testing set, $\nu(\mu(X), \bar{X})$. It is this fact that made it possible [18] to separate and compare numerically the four basic factors responsible for the overestimation of the classical VC bounds. In general, the weak axiom more clearly illustrates the relation between theoretical bounds and empirical methods such as permutation tests, bootstrap, and cross-validation.

Third, if necessary, bounds of the form $Q_\varepsilon(\mu, \mathbb{X}) \le \eta(\varepsilon)$ can easily be carried over from the weak axiom to the strong (Kolmogorov) axiom. To this end, one makes an additional assumption that the objects in $\mathbb{X}$ are chosen randomly and independently of a certain unknown probability distribution. Then, it suffices to take the expectation of both sides of the inequality over the full sample $\mathbb{X}$:

$$\mathrm{P}_{\mathbb{X}}\{\delta_\mu(X) \ge \varepsilon\} = \mathrm{E}_{\mathbb{X}} Q_\varepsilon(\mu, \mathbb{X}) \le \mathrm{E}_{\mathbb{X}}\eta(\varepsilon).$$

If the bound $\eta(\varepsilon)$ does not depend on the full sample $\mathbb{X}$, then it is directly carried over from the weak axiom to the strong axiom. If the bound depends on a certain function of the full sample, $T(\mathbb{X})$, then one should either interpret the value of this function as a priori knowledge or estimate it by the observed part of the sample. In either case the form of the bound remains unchanged under transition from the weak to the strong axiom. Therefore, it is quite feasible to remain within the weak axiom.

2. VAPNIK–CHERVONENKIS BOUNDS AND THEIR IMPROVEMENTS

First, consider the simplest case when the algorithm μ constructs the same classifier $a = \mu(X)$ by any sample $X \subset \mathbb{X}$. Actually, this means that there is no learning. For a fixed classifier a, one estimates the difference between the error rates of this classifier on the hidden and observed samples [18].

Theorem 2.1. *Suppose that a classifier a makes m errors on a full sample*: $n(a, \mathbb{X}) = m$. *Then the following exact bound holds for any* $\varepsilon \in [0, 1]$:

$$\mathrm{P}[\delta(a, X) \ge \varepsilon] = \sum_{s = s_0}^{s_1(\varepsilon)} h_L^{l,m}(s) \equiv H_L^{l,m}(s_1(\varepsilon)). \quad (2.1)$$

Here $h_L^{l,m}(s) = C_m^s C_{L-m}^{l-s} / C_L^l$ *is a hypergeometric distribution*, $s_0 = \max\{0, m - k\}$, *and* $s_1(\varepsilon) = \left\lfloor \frac{l}{L}(m - \varepsilon k) \right\rfloor$.

When $l, k \to \infty$, the right-hand side of (2.1) tends to zero. Therefore, Theorem 2.1 can be considered as an analog of the law of large numbers under the weak axiom. Moreover, the well-known Chernoff, Bennet, Hoeffding, and other bounds [12] can be considered as asymptotic inflated estimates for the *exact* equality (2.1).

To generalize Theorem 2.1 to the case of an arbitrary learning algorithm μ, one should introduce a few more concepts.

The *error vector* of a classifier a on a full sample $\mathbb{X}$ is defined as an L-dimensional binary vector $(a)_{\mathbb{X}} = (I(a, x_i))_{i=1}^{L}$. Since we will mainly deal with the error vectors of classifiers rather than the classifiers themselves, we will use, for short, the symbol a instead of $(a)_{\mathbb{X}}$ and say "vector a."

A *shatter coefficient* of the set of classifiers $\mathbb{A}$ on the sample $\mathbb{X}$ is the number of different error vectors $(a)_{\mathbb{X}}$ generated by all possible classifiers $a \in \mathbb{A}$.

Denote by A the set of error vectors generated by classifiers of the form $a = \mu X$ on all possible training subsamples X:

$$A = \{(\mu X)_{\mathbb{X}}: X \in \mathbb{X}_L^l\}.$$

Note that the cardinality of the set of classifiers $\{\mu X: X \in \mathbb{X}_L^l\}$ is no greater than C_L^l. It may even be strictly less than C_L^l because the algorithm μ may construct identical classifiers from different samples. The shatter coefficient $|A|$ may be still less because different algorithms may generate identical error vectors. In the general case, $|A| \le C_L^l$.

The set of error vectors A is partitioned into $L + 1$ disjoint subsets $A = A_0 \cup \ldots \cup A_L$, where $A_m = \{a \in A: n(a, \mathbb{X}) = m\}$ is the set of vectors with m errors. We will say that A is split into error levels.

A sequence of shatter coefficients $|A_m|$, $m = 0, \ldots, L$, is called a *shatter profile* of the set of classifiers $\mathbb{A}$ on the sample $\mathbb{X}$ [18].

To obtain upper bounds for the probability of overfitting that are valid for any algorithm μ, in the VC theory [16, 15] and in a number of subsequent works (see the surveys [3, 1]), the *uniform convergence principle* was introduced. The functional Q_ε is replaced by its upper bound $\tilde{Q}_\varepsilon$ —the probability of large deviation of rates in two subsamples:

$$Q_\varepsilon \le \tilde{Q}_\varepsilon = \mathrm{P}[\max_{a \in A} \delta(a, X) \ge \varepsilon]. \quad (3.1)$$

In the original papers [16, 15], the authors used a still looser bound: the maximum was taken over all classifiers of the original set $\mathbb{A}$.

Theorem 2.2. *If an algorithm μ minimizes the empirical risk and all vectors $a \in A$ have the same error rate* $m = n(a, \mathbb{X})$, *then the upper bound* (2.2) *turns into the exact equality* $Q_\varepsilon = \tilde{Q}_\varepsilon$.

Proof. The minimization of the empirical risk $\nu(a, X) = \frac{s}{l}$ for a fixed m is equivalent to the minimization of the overfitting $\delta(a, X) = \frac{m-s}{k} - \frac{s}{l} = \frac{ml - sL}{lk}$.

If the set A is split into error levels, then inequality (2.2) becomes an overestimated upper bound because the overfitting attains its maximum for classifiers a that are characterized not only by small $s = n(a, X)$ but also by large $m = n(a, \mathbb{X})$. In practice one almost always meets the phenomenon of splitting. This fact is due to the universal character of the sets of classifiers $\mathbb{A}$ used. For every specific problem defined by an error indicator I and a sample $\mathbb{X}$, only a small part of classifiers of the set have a low error level. The overwhelming majority of classifiers are intended for different problems and make up about half of the errors in a given problem. Experiments confirm that the distribution of classifiers by error levels $m = 0, \ldots, L$ has the form of a narrow peak concentrated near the worst level $m = L/2$ [11, 10].

Thus, the requirement of uniform convergence is too strong. It only gives a sufficient condition for learnability.

An attempt to take into account splitting within the weak axiom was made in [18], where a bound was obtained that depends on the shatter profile $|A_m|_{m=0}^{L}$ rather than on the shatter coefficient $|A|$. Below, we present a shorter proof of the same bound. Here this bound is derived by the uniform convergence principle. With regard to Theorem 2.2, this means that this bound only partially takes into account splitting, although it depends on the shatter profile.

Theorem 2.3. *The following bounds are valid for any* μ, $\mathbb{X}$, *and* $\varepsilon \in [0, 1)$:

$$Q_\varepsilon \le \sum_{m=0}^{L} |A_m| H_L^{l,m}(s_1(\varepsilon)) \qquad (3.1)$$

$$\le |A| \max_{m=1,\ldots,L} H_L^{l,m}(s_1(\varepsilon)). \qquad (3.1)$$

Proof. Let us show that these bounds are valid for the functional $\tilde{Q}_\varepsilon$. We estimate the maximum of the quantities $[\delta(a, X) \ge \varepsilon]$ by their sum (the union bound) and apply splitting into error levels $|A| = + \ldots + |A_L|$:

$$\tilde{Q}_\varepsilon = \mathrm{P}[\max_{a \in A} \delta(a, X) \ge \varepsilon] = \mathrm{P}\max_{a \in A}[\delta(a, X) \ge \varepsilon]$$

$$\le \sum_{a \in A} \mathrm{P}[\delta(a, X) \ge \varepsilon] = \sum_{m=0}^{L} \sum_{a \in A_m} \mathrm{P}[\delta(a, X) \ge \varepsilon]$$

$$= \sum_{m=0}^{L} |A_m| H_L^{l,m}(s_1(\varepsilon)) \le |A| \max_m H_L^{l,m}(s_1(\varepsilon)).$$

The empirical analysis of overestimation factors of the bound (2.4) has shown that two factors are most important [18]. The first is the neglect of splitting; it leads to overestimation of the bound by a factor of 10^3–10^5. The second is the neglect of the similarity of classifiers; it leads to overestimation by a factor of 10^3–10^4. Other factors are of a technical character and are rather easily removed; they give overestimation by a factor of 10^1–10^2 in total. In particular, the third factor of overestimation, which is associated with the replacement of the shatter profile $|A_m|$ by a single scalar shatter coefficient $|A|$ (transition from (2.3) to (2.4)) turned out to be not as essential as one could expect.

The effect of splitting and the related shell bounds were studied by Langford [10, 11]. Unfortunately, these bounds have some drawbacks. First, they are too cumbersome both in form and for computation. To estimate the shatter profile, one should generate a random subset of classifiers from $\mathbb{A}$ by the Monte Carlo method. Second, these bounds do not give a crucial gain in accuracy compared with the classical VC bounds. Another approach is based on the *algorithmic luckiness function*, which orders all the classifiers from the set by their preference with respect to a given sample; then, following the classical VC theory, the union bound is applied to estimate covering numbers, thus again leading to overestimation [7].

The second factor of overestimation—the neglect of the similarity of classifiers—arises from the union bound. The overestimation of the union bound is the higher the more similar the error vectors of the classifiers are. The effect of the similarity of classifiers on the probability of overfitting has hardly been studied, except for the studies by Bax [2] and Sill [14], in which no significant improvements of the bounds were obtained.

Following Bax [2], one can refine Theorem 2.3. If the set of error vectors A is clustered using the Hamming distance into $S(r)$ clusters of radius r each, then

$$\mathrm{P}\left[\delta_\mu(X) \ge \varepsilon + \frac{r}{l}\right] \le S(r) \max_m H_L^{l,m}(s_1(\varepsilon)).$$

In particular, Bax showed that if the set $\mathbb{A}$ is linear in its parameters, then $S(r) \le \frac{1}{2r+1}|A|$. Unfortunately, this bound remains strongly overestimated even after optimization with respect to r.

Sill [13, 14] considered the parametric sets of classifiers $\mathbb{A} = \{a(x, \gamma) : \gamma \in \mathbb{R}^d\}$ that possess the property of *connectedness*, which consists in the following. Under a continuous variation of any of the coordinates of the parameter vector γ, a variation of the error vector of the classifier a(x, γ) occurs only on a single object. One can show (under some technical assumptions) that a simultaneous variation of several coordinates has zero probability. Owing to this property, the set of error vectors of all classifiers of the set almost always forms a connected graph whose edges correspond to pairs of vectors that differ only on a single object. The property of connectedness is inherent in many classifiers having a separating surface continuous in the parameters: linear classifiers, support vector machines with continuous kernels, neural networks with continuous activation functions, decision trees with threshold stamps, and many others. Rewriting the expressions from [13, 14] in the weak axiom, one can easily show that a connected set $\mathbb{A}$ satisfies the following bound:

$$\mathsf{P}[\delta_\mu(X) \ge \varepsilon] \le \frac{1}{\sqrt{\pi L}}|A|\max_m H_L^{l,m}(s_1(\varepsilon)),$$

which differs from (2.4) only by a factor $\sqrt{\pi L}$, which is much less than the degree of overestimation and, hence, does not give a considerable improvement in accuracy.

The question arises: Why have many attempts failed although efforts have been made to take into account the effects of splitting and similarity, which are the basic factors of overestimation?

3. A FAMILY OF TWO CLASSIFIERS

The aim of this section is twofold. First, we show that it is possible in principle to obtain *exact* bounds for the probability of overfitting on the basis of only the weak probability axiom and simple combinatorial arguments. Second, we show that the overfitting arises even in the simplest case, and the effects of splitting and similarity reduce the probability of overfitting.

Consider a set of two classifiers, $\mathbb{A} = \{a_1, a_2\}$. Take, as μ, the *empirical risk minimization* algorithm. When the choice of the best classifier on the training sample is ambiguous, i.e., when $\nu(a_1, X) = \nu(a_2, X)$, we will take the worst case, assuming that the classifier chosen is that with the larger number of errors on the full sample.

Theorem 3.1. *Suppose that, in a sample* $\mathbb{X}$, *there are* m_0 *objects on which both classifiers make an error*, m_1 *objects on which only* a_1 *makes an error,* m_2 *objects on which only* a_2 *makes an error, and* m_3 *objects on which neither classifier makes an error. Let, for definiteness,* $m_1 \le m_2$:

$$a_1 = (1, \ldots, 1, 1, \ldots, 1, 0, \ldots, 0, 0, \ldots, 0),$$
$$a_2 = (\underbrace{1, \ldots, 1}_{m_0}, \underbrace{0, \ldots, 0}_{m_1}, \underbrace{1, \ldots, 1}_{m_2}, \underbrace{0, \ldots, 0}_{m_3}).$$

Then the following exact bound holds for any $\varepsilon \in [0, 1)$:

$$Q_\varepsilon = \sum_{s_0=0}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s_2=0}^{m_2}\sum_{s_3=0}^{m_3}\frac{C_{m_0}^{s_0}C_{m_1}^{s_1}C_{m_2}^{s_2}C_{m_3}^{s_3}}{C_L^l}$$
$$\times [s_0 + s_1 + s_2 + s_3 = l]$$

$$\times \left([s_1 < s_2]\left[s_0 + s_1 \le \frac{l}{L}(m_0 + m_1 - \varepsilon k)\right]\right.$$
$$\left. + [s_1 \ge s_2]\left[s_0 + s_2 \le \frac{l}{L}(m_0 + m_2 - \varepsilon k)\right]\right).$$

Proof. The empirical risk minimization algorithm chooses the classifier a_1 when $\nu(a_1, X) < \nu(a_2, X)$ and the classifier a_2 otherwise. Hence,

$$Q_\varepsilon = \frac{1}{C_L^l}\sum_{X \in \mathbb{X}_L^l}[\nu(a_1, X) < \nu(a_2, X)]$$
$$\times [\nu(a_1, \bar{X}) - \nu(a_1, X) \ge \varepsilon]$$
$$+ \frac{1}{C_L^l}\sum_{X \in \mathbb{X}_L^l}[\nu(a_1, X) \ge \nu(a_2, X)]$$
$$\times [\nu(a_2, \bar{X}) - \nu(a_2, X) \ge \varepsilon].$$

Divide the set $\mathbb{X}$ into four subsets: X_0, a subset of objects on which both classifiers make an error; X_1, a subset of objects on which only a_1 makes an error; X_2, a subset of objects on which only a_2 makes an error; and X_3, a subset of all the other objects. Obviously, $m_i = |X_i|$. Denote by $s_i = |X_i \cap X|$ the set of objects from X_i that belongs to the training sample.

In this notation, the error rates of the classifiers a_1 and a_2 on the samples X and $\bar{X}$ are given by

$$\nu(a_1, X) = \frac{s_0 + s_1}{l}, \quad \nu(a_1, \bar{X}) = \frac{m_0 + m_1 - s_0 - s_1}{k},$$
$$\nu(a_2, X) = \frac{s_0 + s_2}{l}, \quad \nu(a_2, \bar{X}) = \frac{m_0 + m_2 - s_0 - s_2}{k}.$$

The number of partitions under which the set of values (s_0, s_1, s_2, s_3) is realized is given by

$$\sum_{X \in \mathbb{X}_L^l}\prod_{i=0}^{3}[s_i = |X_i \cap X|] = C_{m_0}^{s_0}C_{m_1}^{s_1}C_{m_2}^{s_2}C_{m_3}^{s_3}. \qquad (3.1)$$

Hence, s_0, s_1, s_2, and s_3 must satisfy the following constraints:

$$0 \le s_0 \le m_0, \quad 0 \le s_1 \le m_1,$$
$$0 \le s_2 \le m_2, \quad 0 \le s_3 \le m_3.$$

In addition, s_0, s_1, s_2, and s_3 must satisfy the relation $s_0 + s_1 + s_2 + s_3 = l$.

Thus,

$$Q_\varepsilon = \frac{1}{C_L^l}\sum_{X \in \mathbb{X}_L^l}\sum_{s_0=0}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s_2=0}^{m_2}\sum_{s_3=0}^{m_3}[s_0 + s_1 + s_2 + s_3 = l]$$
$$\times \prod_{i=0}^{3}[s_i = |X_i \cap X|]$$

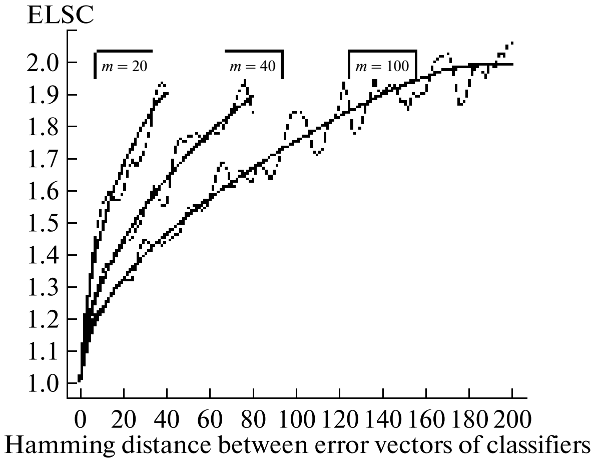

Fig. 1. Upper bound for the ELSC $\bar{\Delta}$ as a function of the diversity of classifiers when the classifiers make the same number of errors, $m_1 = m_2$. The three graphs correspond to three different values of the number of errors on the full sample: $m = n(a_i, \mathbb{X}) = m_1 + m_0 \in \{20, 40, 100\}$.

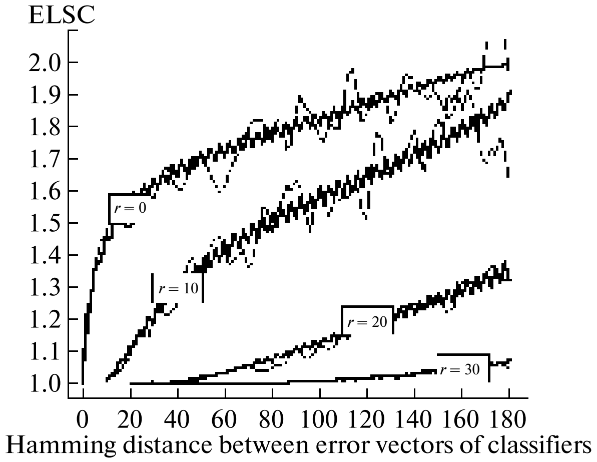

Fig. 2. Upper bound for the ELSC $\bar{\Delta}$ as a function of the diversity of classifiers when $m_0 = 20$ and the number of errors made by the second classifier is greater by r, $m_2 = m_1 + r$. The four graphs correspond to four different values of $r \in \{0, 10, 20, 30\}$.

$$\times\Big([s_1 < s_2]\Big[\frac{m_0 + m_1 - s_0 - s_1}{k} - \frac{s_0 + s_1}{l} \ge \varepsilon\Big]$$

$$+ [s_1 \ge s_2]\Big[\frac{m_0 + m_2 - s_0 - s_2}{k} - \frac{s_0 + s_2}{l} \ge \varepsilon\Big]\Big).$$

Interchanging the summation signs and substituting (3.1) into this formula, we obtain the required exact bound.

Along with the probability of overfitting, in experiments it is convenient to estimate the *effective local shatter coefficient* (ELSC) introduced in [18]. This is the value of the shatter coefficient $|A|$ for which bound (2.4) is not overestimated. Comparing (2.1) and (2.4), we obtain the following expression for the ELSC:

$$\Delta = \frac{\mathrm{P}[\delta_\mu(X) \ge \varepsilon]}{\max\limits_{a \in A} \mathrm{P}[\delta(a, X) \ge \varepsilon]}.$$

In this paper, we estimate the upper bound for the ELSC, which has a more natural interpretation. It shows how much the probability of overfitting of algorithm μ is greater than the probability of a large deviation of error rates for the best classifier in the set:

$$\bar{\Delta} = \frac{\mathrm{P}[\delta_\mu(X) \ge \varepsilon]}{\max\limits_{a \in A} \mathrm{P}[\delta(a, X) \ge \varepsilon]}.$$

It is obvious that, in the case of a two-element set of classifiers, $1 \le \bar{\Delta} \le 2$.

Figures 1 and 2 show the upper bound for the ELSC $\bar{\Delta}$ as a function of the diversity of classifiers for $l = k = 100$ and $\varepsilon = 0.05$. The Hamming distance $\rho(a_1, a_2) = m_1 + m_2$ between error vectors is taken as a natural measure of diversity. Thin solid lines show the ELSC bounds calculated by the Monte Carlo method using 1000 random partitions.

The charts lead to the following conclusions.

1. Overfitting is provoked by the choice of a classifier from an incomplete data sample $X \subset \mathbb{X}$ even though the choice is made merely between two classifiers.

2. If the classifiers make the same number of errors on $\mathbb{X}$ ($m_1 = m_2$) but are maximally different ($m_0 = 0$), then the VC bound $\bar{\Delta} = 2$ is either attained or nearly attained.

3. If the classifiers are similar, then the ELSC approaches 1; i.e., from the viewpoint of overfitting, two similar classifiers behave almost as a single classifier.

4. If the classifiers are different in the number of errors, $r = m_2 - m_1 > 0$, then the VC bound is not attained either. The greater r, the lower the probability of overfitting.

The main conclusion is as follows: the effect of overfitting arises even in the simplest case when there are only two classifiers. In this case, the properties of splitting and similarity reduce the probability of overfitting.

4. EXPERIMENTS WITH CHAINS OF CLASSIFIERS

The aim of this section is to demonstrate, by a specific example, that computationally tight bounds for the probability of overfitting can only be obtained by simultaneous consideration of the splitting of a set of classifiers and the similarity of classifiers within a set.

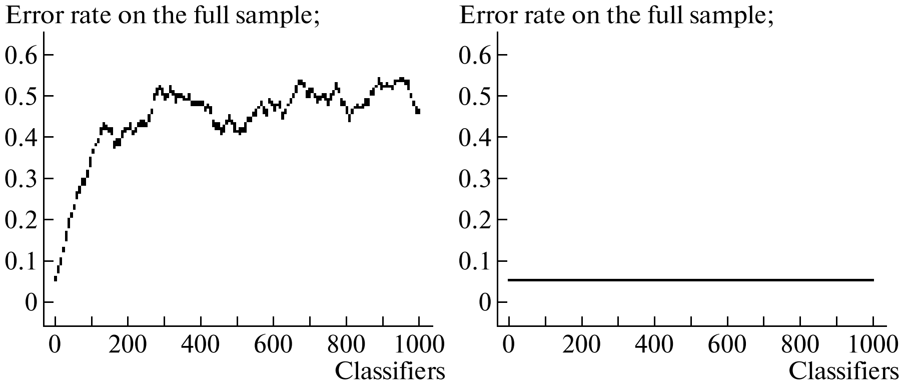

Fig. 3. Chains with and without splitting. The function $\nu(a_t, \mathbb{X})$ versus t for $l = k = 100$ and $m = 10$.

A sequence of classifiers $\{a_1, ..., a_D\}$ is called a *chain* if the Hamming distance between the error vectors a_{t-1} and a_t is 1 for any $t = 2, ..., D$. A chain is the simplest example of a connected set of classifiers [14].

The probability of overfitting as a function of the sequence length D was investigated experimentally. To this end, we constructed two types of model chains that were defined directly by a sequence of error vectors $a_1, ..., a_D$.

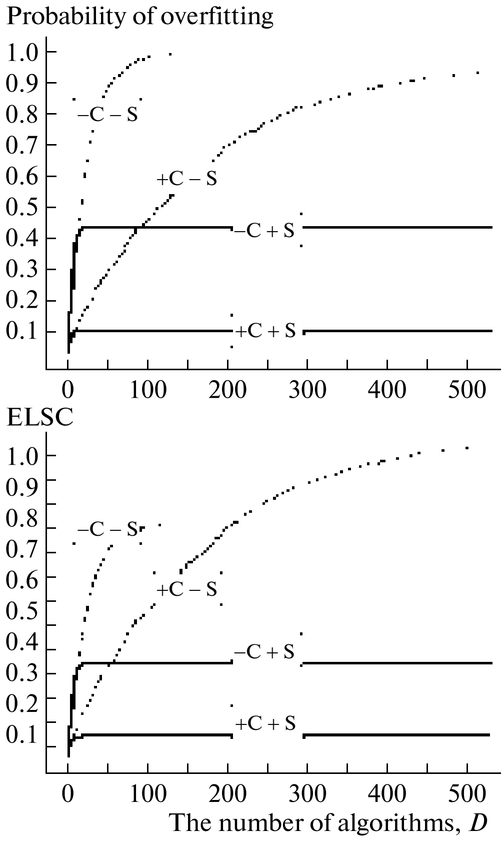

Fig. 4. Probability of overfitting Q_ε and the ELSC $\bar{\Delta}$ as a function of the number D of classifiers (a "simple problem": the error rate of the best algorithm is $\nu(a_1, \mathbb{X}) = 0.05$).

1. A *split chain* (Fig. 3, left). The best algorithm a_1 makes m errors on the full sample. Each subsequent error vector a_t is obtained from a_{t-1} by inverting one randomly chosen coordinate. If a chain is sufficiently long ($D \gg L$), then most classifiers make m errors, which is close to $L/2$.

2. A *nonsplit chain* (Fig. 3, right). The number of errors of classifiers on the full sample alternates between two values m and $m + 1$.

For each chain, we constructed a corresponding *nonchain* $\{a'_t, ..., a'_D\}$, which consists of essentially different classifiers. The error vectors a'_t were generated randomly but so that $\nu(a'_t, \mathbb{X}) = \nu(a_t, \mathbb{X})$ for all $t = 1, ..., D$. Thus, the neighboring classifiers a_{t-1} and a_t in nonchains were not similar.

In total, we constructed four finite sets of classifiers with identical values of the parameters D and m. The juxtaposition of these four cases has allowed us to distinguish between the effects of *similarity* (a chain or a nonchain) and *splitting* (m errors made either by all the classifiers or only by the best one) on the probability of overfitting.

Figures 4 and 5 show the probability of overfitting Q_ε and ELSC $\bar{\Delta}$ as a function of the number D of classifiers for four types of sets for $l = k = 100$ and $\varepsilon = 0.05$. The probabilities Q_ε were calculated by the Monte Carlo method using 1000 random partitions. The following notation is used in the figures: $+C$ denotes the chain, $-C$ denotes the nonchain, $+S$ denotes the split, and $-S$ denotes the nonsplit.

These results lead to the following conclusions.

1. The dependence of the ELSC $\bar{\Delta}$ on the number D of classifiers shows what value the shatter coefficient should take, instead of D, so that bound (2.4) would not be overestimated. This value may prove to be much

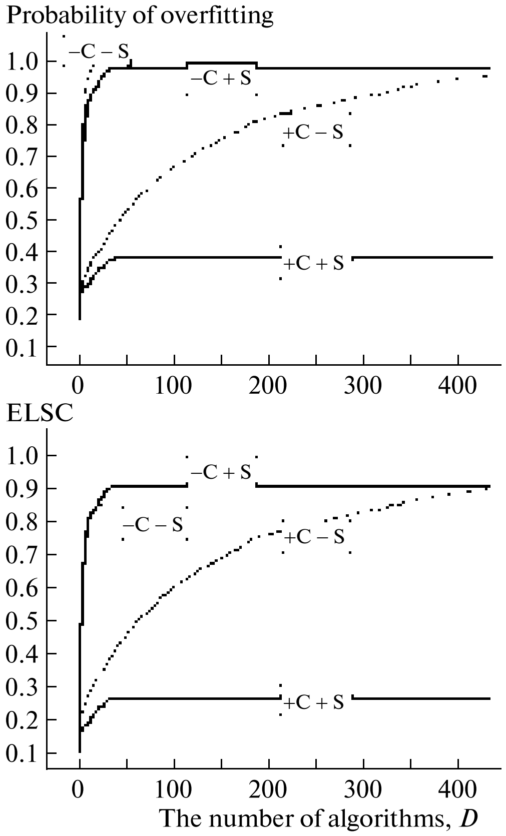

Fig. 5. Probability of overfitting Q_ε and the ELSC $\bar{\Delta}$ as a function of the number D of classifiers (a "hard problem": the error rate of the best algorithm is $\nu(a_1, \mathbb{X}) = 0.25$).

less than D. At some instant, the probability of overfitting reaches a certain maximal value $Q_{\max}$; in this case, the ELSC reaches a horizontal asymptote and becomes independent of D. At the same time, the VC bound is linear in D and has no horizontal asymptote. The VC bound is attained only for nonchains and only for small D (in this experiment, for $D < 8$).

2. For chains, the probability of overfitting Q_ε grows much more slowly as the number D of classifiers increases. Thus, owing to the connectedness, the number of classifiers in a set can be much greater than that predicted by the VC theory.

3. For splits (thick solid lines in the figures), the probability of overfitting Q_ε may not reach 1 even for very large D. At the same time, for nonsplit chains, $Q_{\max}$ reaches the value 1 for D on the order of hundreds. Thus, it is the splitting property that reduces the horizontal asymptote of $Q_{\max}$ to a level much below unity. Note that this phenomenon cannot be explained on the basis of the uniform convergence principle, because, according to Theorem 2.2, $Q_\varepsilon = \tilde{Q}_\varepsilon$ only in the absence of splitting.

4. For relatively simple problems, when a low-error classifier exists, the presence of splitting strongly reduces the probability of overfitting compared with the absence of splitting (Fig. 4). As the complexity of a problem increases, the effect of splitting decreases (Fig. 5).

5. For large D, only the presence of both a chain and splitting considerably reduces the probability of overfitting (the lower curves in the figures). The fact that precisely this case is widespread in practice gives grounds for optimism.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 08-07-00422) and by the Program "Algebraic and Combinatorial Methods in Mathematical Cybernetics" of the Department of Mathematics, Russian Academy of Sciences.

REFERENCES

1. K. V. Vorontsov, "A Survey on Modern Research," *Tavrich. Vestn. Inf. Mat.,* No. 1, 5–24 (2004).
2. E. T. Bax, "Similar Classifiers and VC Error Bounds," Tech. Rep. CalTech-CS-TR97-14: 6 1997.
3. S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of Classification: A Survey of Some Recent Advances," *ESIAM: Probab. Stat.,* No. 9, 323–375 (2005).
4. B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (SIAM, Philadelphia, 1982).
5. R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics* (Addison-Wesley, Reading, 1994), p. 657.
6. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001).
7. R. Herbrich and R. Williamson, "Algorithmic Luckiness," *J. Machine Learning Res.*, No. 3, 175–212 (2002).
8. M. J. Kearns, Y. Mansour, A. Y. Ng, and D. Ron, "An Experimental and Theoretical Comparison of Model Selection Methods," in *Proceedings of the 8th Conference on Computational Learning Theory, Santa Cruz, California, US, 1995*, pp. 21–30.
9. R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Palais de Congress Montreal, Quebec, Canada, 1995*, pp. 1137–1145.
10. J. Langford, "Quantitatively Tight Sample Complexity Bounds," Ph.D. Thesis (Carnegie Mellon Thesis, 2002).
11. J. Langford and D. McAllester, "Computable Shell Decomposition Bounds," in *Proceedings of the 13th Annual Conference on Computer Learning Theory* (Morgan Kaufmann, San Francisco, CA, 2000), pp. 25–34.

12. G. Lugosi, “On Concentration-of-Measure Inequalities,” in *Machine Learning Summer School* (Australian National University, Canberra, 2003).

13. J. Sill, “Generalization Bounds for Connected Function Classes,” citeseer.ist.psu.edu/127284.html.

14. J. Sill, “Monotonicity and Connectedness in Learning Systems,” Ph.D. Thesis (California Inst. Technol., 1998).

15. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

16. V. Vapnik and A. Chervonenkis, “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities,” *Theory Probab. Its Appl.* **16** (2), 264–280 (1971).

17. K. V. Vorontsov, “Combinatorial Substantiation of Learning Algorithms,” *Comput. Math. Math. Phys.* **44** (11), 1997–2009 (2004).

18. K. V. Vorontsov, “Combinatorial Probability and the Tightness of Generalization Bounds,” *Pattern Recognit. Image Anal.* **18** (2), 243–259 (2008).

Konstantin Vorontsov was born in 1971. He graduated from the Faculty of Applied Mathematics and Control, Moscow Institute of Physics and Technology, in 1994. He received his candidate’s degree in 1999. Currently he is with the Dorodnitsyn Computing Centre, Russian Academy of Sciences. His scientific interests include statistical learning theory, machine learning, data mining, probability theory, and combinatorics. He is the author of 40 papers. Homepage: www.ccas.ru/voron.