ISSN 1064-5624, Doklady Mathematics, 2009, Vol. 80, No. 3, pp. 793–796. © Pleiades Publishing, Ltd., 2009.
Original Russian Text © K.V. Vorontsov, 2009, published in Doklady Akademii Nauk, 2009, Vol. 429, No. 1, pp. 15–18.

MATHEMATICS

Tight Bounds for the Probability of Overfitting

K. V. Vorontsov

Presented by Academician Yu.I. Zhuravlev April 7, 2009

Received April 21, 2009

DOI: 10.1134/S1064562409060032

The derivation of tight generalization bounds has remained an open problem in the statistical learning theory starting from the works by Vapnik and Chervonenkis [1, 2]. Numerous attempts were made to improve their results (see [3, 4]), but the best of the known bounds are still highly overestimated [5, 6] and are not always suitable for the control of a learning process. Another open problem is whether overfitting is related to still unexamined subtler properties of statistical learning methods. In this paper, a combinatorial approach is developed that leads to tight bounds for the probability of overfitting in a number of special cases.

Suppose that we are given a finite set of objects $\mathbb{X} = \{x_1, x_2, ..., x_L\}$, which is called a general sample, and a set A whose elements are called algorithms. There exists a binary function I: $A \times \mathbb{X} \to \{0, 1\}$, which is called an error indicator. If $I(a, x) = 1$, then we say that the algorithm a makes an error on the object x. The error vector of a is the L-dimensional binary vector $(I(a, x_i))_{i=1}^{L}$. The number of errors of a on a sample $X \subseteq \mathbb{X}$ is defined as

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

and the frequency of errors or the empirical risk of a on X is defined as

$$\nu(a, X) = \frac{1}{|X|} n(a, X).$$

Let $\ell < L$ be a fixed positive integer. Denote by $[\mathbb{X}]^\ell$ the set of all ℓ-element subsets of $\mathbb{X}$. Obviously, its cardinality is C_L^ℓ.

A learning method is a mapping μ: $[\mathbb{X}]^\ell \to A$ that an arbitrary training set $X \in [\mathbb{X}]^\ell$ transforms into an algorithm $a = \mu(X)$ from A. A learning method μ is called an empirical risk minimization method if

$$\mu(X) = \operatorname{argmin}_{a \in A} n(a, X). \quad (1)$$

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

The deviation of the error frequency of a on X from that on $\bar{X} = \mathbb{X} \backslash X$ is defined as $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

The overfitting of a method μ on a set X is the error frequency deviation for the algorithm $a = \mu(X)$:

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

A method μ is said to be overfit on a set X if $\delta_\mu(X) \geq \varepsilon$ for a given $\varepsilon \in (0, 1)$.

Following the weak probability assumptions [7], we assume that all C_L^ℓ partitions of X into an observed training set X of length ℓ and a hidden test set $\bar{X}$ of length $k = L - \ell$ are realized with an identical probability. The goal of this work is to derive tight bounds for the probability of overfitting for μ:

$$Q_\varepsilon \equiv \mathrm{P}[\delta_\mu(X) \geq \varepsilon] = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} [\delta_\mu(X) \geq \varepsilon]. \quad (2)$$

Here and below, the logical expression in square brackets means [true] = 1 and [false] = 0, respectively.

For a fixed algorithm a that makes $m = n(a, \mathbb{X})$ errors on the general sample, the probability of making exactly s errors on X is described by the hypergeometric probability function

$$\mathrm{P}[n(a, X) = s] = h_L^{\ell, m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell},$$

where $m \in \{0, 1, ..., L\}$ and the argument s takes integer values from $s_0 = \max\{0, m - k\}$ to $s_1 = \min\{m, \ell\}$. For all the other integers m and s, the binomial coefficients C_m^s and the function $h_L^{\ell, m}(s)$ are extended by zero. The probability of a large error frequency deviation for a is described by the hypergeometric distribution function

$$\mathrm{P}[\delta(a, X) \geq \varepsilon] = \mathrm{P}[n(a, X) \leq s] = H_L^{\ell, m}(s) = \sum_{s'=s_0}^{\lfloor s \rfloor} h_L^{\ell, m}(s'), \quad (3)$$

where $s = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$ is the largest value of $n(a, X)$ at which $\delta(a, X) = \frac{m-s}{k} - \frac{s}{\ell} \geq \varepsilon$. As $\ell, k \to \infty$, the right-hand side of (3) tends to zero and is a tight bound for the convergence rate of the frequencies on two sets.

The Vapnik–Chervonenkis classical bound [8] is also easy to restate under the weak probability assumptions [4, 7]:

$$Q_\varepsilon \leq \Delta \max_{m = 0, \ldots, L} H_L^{\ell, m}\left(\frac{\ell}{L}(m - \varepsilon k)\right), \qquad (4)$$

where Δ is the diversity coefficient of A, which is equal to the number of different error vectors generated by all possible algorithms a from A. An experimental analysis of major causes of overestimated bound (4) shows that the probability of overfitting depends substantially not only on the number of different error vectors but also on the degree of their difference [7]. To derive tighter bounds, we have to take into account the splitting and similarity of algorithms in A.

Splitting effect. In practical situations, A generally splits according to the error frequency $\nu(a, \mathbb{X})$, and most of the algorithms concentrate within the worst-frequency region (about 50%). Only a small fraction of algorithms have high chances to be chosen by an empirical risk minimization method. Experiments [7] with real-life classification problems show that the bound for Q_ε can degrade by 10^2–10^5 times if the splitting effect is neglected.

Similarity effect. The set A may contain a large number of pairs of similar algorithms. Specifically, most classification algorithms used in practice have a separating surface that is continuous with respect to the parameters. Therefore, they have the connectivity property [9]. The set A is called connected with respect to $\mathbb{X}$ if, for any algorithm $a \in A$, there is another algorithm $a' \in A$ such that their error vectors differ only on a single object. Experiments [7] show that the bound for Q_ε can degrade by 10^3–10^4 times if the similarity of algorithms is neglected.

In experiments with chains of algorithms [10], the probability of overfitting is considerably reduced only if splitting and connectivity are both taken into account. If one of them is neglected in deriving a bound for Q_ε, the effect of taking into account the other can be nullified. Attempts to take into account them separately do not lead to radical improvements of bound tightness [5, 6, 9, 11].

In this paper, tight bounds for the probability of overfitting are presented that are based on the assumption that, for each algorithm $a \in A$, the conditions under which $\mu(X) = a$ can be explicitly written. Assume that A is a finite set and all the algorithms have pairwise distinct error vectors.

Conjecture 1. *Let A, $\mathbb{X}$, and μ be such that, for each algorithm $a \in A$, a pair of subsets $X_a \subset \mathbb{X}$ and $X'_a \subset \mathbb{X}$ can be found such that, for any $X \in [\mathbb{X}]^\ell$,*

$$\mu(X) = a \Leftrightarrow (X_a \subseteq X)\ u\ (X'_a \subseteq \bar{X}).$$

The objects in X_a are call generating (reference), while the objects in X'_a re called destroying (noise). The remaining objects are referred to as neutral for a. For each $a \in A$, we introduce the following notation:

$L_a = L - |X_a| - |X'_a|$ is the number of neutral objects;

$\ell_a = \ell - |X_a|$ is the number of neutral training objects;

$m_a = n(a, \mathbb{X}) - n(a, X_a) - n(a, X'_a)$ is the number of errors on neutral objects;

$s_a(\varepsilon) = \frac{\ell}{L}(n(a, \mathbb{X}) - \varepsilon k) - n(a, X_a)$ is the largest number of errors on neutral training objects for which $\delta(a, X) \geq \varepsilon$.

Theorem 1. *If Conjecture 1 holds, then, for any $\varepsilon \in (0, 1)$,*

$$Q_\varepsilon = \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}(s_a(\varepsilon)), \quad P_a = \mathrm{P}[\mu(X) = a] = \frac{\mathrm{C}_{L_a}^{\ell_a}}{\mathrm{C}_L^\ell}.$$

Conjecture 1 and Theorem 1 yield the following generalization.

Conjecture 2. *Let A, $\mathbb{X}$, and μ be such that, for each algorithm $a \in A$, there is a finite index set V_a and, for each index $v \in V_a$, there are subsets $X_{av} \subset \mathbb{X}$ and $X'_{av} \subset \mathbb{X}$ and coefficients $c_{av} \in \mathbb{R}$ such that, for any $X \in [\mathbb{X}]^\ell$*

$$\mu(X) = a \Leftrightarrow \sum_{v \in V_a} c_{av}[X_{av} \subseteq X][X'_{av} \subseteq \bar{X}] = 1.$$

Specifically, if all $c_{av} = 1$, then this condition means that the same algorithm a is produced by training under several various selection methods for generating and destroying objects.

For each $a \in A$ and $v \in V_a$, we introduce the following notation:

$$L_{av} = L - |X_{av}| - |X'_{av}|,$$
$$\ell_{av} = \ell - |X_{av}|,$$
$$m_{av} = n(a, \mathbb{X}) - n(a, X_{av}) - n(a, X'_{av}),$$
$$s_{av}(\varepsilon) = \frac{\ell}{L}(n(a, \mathbb{X}) - \varepsilon k) - n(a, X_{av}).$$

Theorem 2. *If Conjecture 2 holds, then the probability of an algorithm a obtained by training is*

$$\mathrm{P}[\mu(X) = a] = \sum_{v \in V_a} c_{av} P_{av}, \quad P_{av} = \frac{\mathrm{C}_{L_{av}}^{\ell_{av}}}{\mathrm{C}_L^\ell};$$

and the probability of overfitting is

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H_{L_{av}}^{\ell_{av}, m_{av}}(s_{av}(\varepsilon)).$$

In contrast to Conjecture 1, Conjecture 2 holds under rather weak assumptions made about $\mathbb{X}$, A, and μ.

Theorem 3. *$A = \{a_1, a_2, \ldots, a_D\}$ and all the algorithms have pairwise distinct error vectors. Let μ be an empirical risk minimization method such that, if the minimum in (1) is reached on several algorithms in A, then μ chooses an algorithm with larger $n(a, \mathbb{X})$ and, if the maximum of $n(a, \mathbb{X})$ is reached on several algorithms, then an algorithm with a smaller number is picked.*

Then Conjecture 2 holds.

In what follows, it is assumed that μ satisfies the conditions of Theorem 3. Consider four special cases for which tight bounds for the probability of overfitting are obtained using Theorems 1 and 2.

The Hamming distance between the error vectors of algorithms is defined as

$$\rho(a, a') = \sum_{i=1}^{L} \left| I(a, x_i) - I(a', x_i) \right|, \quad \forall a, a' \in A.$$

Definition 1. A set of algorithms $a_0, a_1, \ldots, a_D$ is called a chain if $\rho(a_{d-1}, a_d) = 1$, $d = 1, 2, \ldots, D$.

Definition 2. A chain of algorithms $a_0, a_1, \ldots, a_D$ is called monotone if $n(a_d, \mathbb{X}) = m + d$ for some $m \geq 0$.

Theorem 4. *Let $a_0, a_1, \ldots, a_D$ be a monotone chain of algorithms; $n(a_0, \mathbb{X}) = m$; and $L \geq m + D$. Then*

$$Q_\varepsilon = \sum_{d=0}^{k} P_d H_{L-d-1}^{\ell-1, m}(s_d(\varepsilon)), \quad P_d = \frac{C_{L-d-1}^{\ell-1}}{C_L^\ell},$$
$$d = 0, 1, \ldots, D;$$

if $D \geq k$ and

$$Q_\varepsilon = \sum_{d=0}^{D-1} P_d H_{L-d-1}^{\ell-1, m}(s_d(\varepsilon)) + P_D H_{L-D}^{\ell, m}(s_d(\varepsilon)),$$

$$P_d = \frac{C_{L-d-1}^{\ell-1}}{C_L^\ell}, \quad d = 0, 1, \ldots, D-1; \quad P_D = \frac{C_{L-D}^{\ell}}{C_L^\ell},$$

if $D < k$. Here, $P_d = P[\mu(X) = a_d]$ and $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$.

Definition 3. A set of algorithms $a_0, a_1, \ldots, a_D, a_1', a_2', \ldots, a_{D'}'$ (with D' not necessarily equal to D) is called a unimodal chain if the left branch $a_0, a_1, \ldots, a_D$ and the right branch $a_0, a_1', \ldots, a_{D'}'$ are monotone chains.

Assume that, if the minimum in (1) is reached on several algorithms with the same number of errors for both training and general sets, then μ chooses an algorithm from the left branch.

Theorem 5. *Let $a_0, a_1, \ldots, a_D, a_1', a_2', \ldots, a_{D'}'$ be a unimodal chain of algorithms, where $k \leq D$, $m = n(a_0, \mathbb{X})$, and $2D + m \leq L$.*

Then the probability of obtaining each of the algorithms is

$$P_0 = P[\mu(X) = a_0] = \frac{C_{L-2}^{\ell-2}}{C_L^\ell};$$

$$P_d = P[\mu(X) = a_d] = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-2}^{\ell-1}}{C_L^\ell},$$
$$d = 1, 2, \ldots, D;$$

$$P_d' = P[\mu(X) = a_d'] = \frac{C_{L-d-1}^{\ell-1} - C_{L-2d-1}^{\ell-1}}{C_L^\ell},$$
$$d = 1, 2, \ldots, D.$$

The probability of overfitting for $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$ is

$$Q_\varepsilon = \frac{C_{L-2}^{\ell-2}}{C_L^\ell} H_{L-2}^{\ell-2, m}(s_0(\varepsilon))$$
$$+ \sum_{d=1}^{k} \left(2\frac{C_{L-d-1}^{\ell-1}}{C_L^\ell} H_{L-d-1}^{\ell-1, m}(s_d(\varepsilon)) \right.$$
$$- \frac{C_{L-2d-2}^{\ell-1}}{C_L^\ell} H_{L-2d-2}^{\ell-1, m}(s_d(\varepsilon))$$
$$\left. - \frac{C_{L-2d-1}^{\ell-1}}{C_L^\ell} H_{L-2d-1}^{\ell-1, m}(s_d(\varepsilon)) \right).$$

Definition 4. A set of algorithms $a_0, a_1, \ldots, a_D$ is called a unit neighborhood of a_0 if their error vectors are pairwise distinct, $n(a_d, \mathbb{X}) = n(a_0, \mathbb{X}) + 1$, and $\rho(a_0, a_d) = 1$ for $d = 1, 2, \ldots, D$.

Theorem 6. *Let $a_0, a_1, \ldots, a_D$ be a unit neighborhood of a_0; $m = n(a_0, \mathbb{X})$; and $L \geq m + D$. Then*

$$Q_\varepsilon = P_0 H_{L-D}^{\ell-D, m}\left(\frac{\ell}{L}(m - \varepsilon k)\right)$$
$$+ \sum_{d=1}^{D} P_d H_{L-d}^{\ell-d+1, m}\left(\frac{\ell}{L}(m + 1 - \varepsilon k)\right),$$
$$P_0 = \frac{C_{L-D}^{k}}{C_L^k}, P_d = \frac{C_{L-d}^{k-1}}{C_L^k}, d = 1, 2, \ldots, D,$$

where P_d is the probability of an algorithm a_d obtained by training.

The last special case is a two-element set $A = \{a_1, a_2\}$. Even in this simplest case, we can see the phenomenon of overfitting and the splitting and similarity effects, which reduce the probability of overfitting [10].

Theorem 7. *Let both algorithms, only a_0 and a_1 make an error on m_0, m_1, and m_2 objects in $\mathbb{X}$, respectively. Then*

$$Q_\varepsilon = \sum_{s_0=0}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s_2=0}^{m_2}\frac{C_{m_0}^{s_0}C_{m_1}^{s_1}C_{m_2}^{s_2}C_{L-m_0-m_1-m_2}^{\ell-s_0-s_1-s_2}}{C_L^\ell}$$
$$\times\left([s_1\le s_2]\left[s_0+s_1\le\frac{\ell}{L}(m_0+m_1-\varepsilon k)\right]\right.$$
$$\left.+[s_1>s_2]\left[s_0+s_2\le\frac{\ell}{L}(m_0+m_1-\varepsilon k)\right]\right).$$

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 and New-Generation Information Systems" of the Department of Mathematical Sciences of the Russian Academy of Sciences.

REFERENCES

1. V. N. Vapnik and A. Ya. Chervonenkis, Teor. Veroyatn. Prim. **16** (2), 264–280 (1971).
2. V. N. Vapnik and A. Ya. Chervonenkis, *Pattern Recognition Theory* (Nauka, Moscow, 1974) [in Russian].
3. N. Vayatis and R. Azencott, Lect. Notes Comput. Sci. **1572**, 230–240 (1999).
4. K. V. Vorontsov, in *Mathematical Issues of Cybernetics* (Fizmatlit, Moscow, 2004), Vol. 13, pp. 5–36 [in Russian].
5. R. Herbrich and R. Williamson, J. Machine Learning Res., No. 3, 175–212 (2002).
6. J. Langford, PhD Thesis (Carnegie Mellon Univ., Pittsburgh, 2002).
7. K. V. Vorontsov, Pattern Recogn. Image Anal. **18**, 243–259 (2008).
8. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
9. J. Sill, PhD Thesis (California Inst. Tech., Pasadena, 1998).
10. K. V. Vorontsov, in *Pattern Recognition and Image Analysis: New Information Technologies (PRIA-9)* (Nizhni Novgorod, 2008), Vol. 2, pp. 303–306.
11. E. T. Bax, "Similar Classifiers and VC Error Bounds," Tech. Rep. CalTech-CS-TR97-14 (California Inst. Tech., Pasadena, 1997).