[Raudys, 2001] Raudys S., Statistical and neural classifiers, Springer, 2001.

[Mirenkova, 2002] S. V. Mirenkova (Nedel'ko). A method for prediction multidimensional heterogeneous time series in class of logical decision functions // Artificial Intelligence, No 2, 2002, p. 197–201. (in Russian).

Author's Information

Svetlana Valeryevna Nedel'ko – Institute of Mathematics SB RAS, Laboratory of Data Analysis, 630090, pr. Koptyuga, 4, Novosibirsk, Russia, e-mail: nedelko@math.nsc.ru

EVALUATING MISCLASSIFICATION PROBABILITY USING EMPIRICAL RISK[1]

Victor Nedel'ko

Abstract: The goal of the paper is to estimate misclassification probability for decision function by training sample. Here are presented results of investigation an empirical risk bias for nearest neighbours, linear and decision tree classifier in comparison with exact bias estimations for a discrete (multinomial) case. This allows to find out how far Vapnik–Chervonenkis risk estimations are off for considered decision function classes and to choose optimal complexity parameters for constructed decision functions. Comparison of linear classifier and decision trees capacities is also performed.

Keywords: pattern recognition, classification, statistical robustness, deciding functions, complexity, capacity, overtraining problem.

ACM Classification Keywords:I.5.1 Pattern Recognition: Statistical Models

Introduction

One of the most important problems in classification is estimating a quality of decision built. As a quality measure, a misclassification probability is usually used. The last value is also known as a risk. There are many methods for estimating a risk: validation set, leave-one-out method etc. But these methods have some disadvantages, for example, the first one decreases a volume of sample available for building a decision function, the second one takes extra computational resources and is unable to estimate risk deviation. So, the most attractive way is to evaluate a decision function quality by the training sample immediately.

But an empirical risk or a rate of misclassified objects from the training sample appears to be a biased risk estimate, because a decision function quality being evaluated by the training sample usually appears much better than its real quality. This fact is known as an overtraining problem.

To solve this problem in [Vapnik, Chervonenkis, 1974] there was introduced a concept of capacity (complexity measure) of a decision rules set. The authors obtained universal decision quality estimations, but these VC–estimations are not accurate and suggest pessimistic risk expectations.

For a case of discrete feature in [Nedel'ko, 2003] there were obtained exact estimations of empirical risk bias. This allows finding out how far VC–estimations are off.

The goal of this paper is to extrapolate the result on continuous case including linear and decision tree classifiers.

[1] The work is supported by RFBR, grant 04-01-00858-a

Formal Problem Definition

A classification task consists in constructing a deciding function that is a correspondence $f : X \to Y$, where X – a features values space and $Y = \{1, k\}$ – a forecasting values space. For simplicity let's assume a number of classes $k = 2$.

For the determination of deciding functions quality one need to assign a loss function: $L : Y^2 \to [0,\infty)$ that for classification task will be $L(y, y') = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$, where $y \in Y,\ y' \in Y$.

By a risk we shall understand an average loss:

$$R(c, f) = \int L(y, f(x)) dP_c[D],$$

where C is a set of probabilistic measures on $D = X \times Y$ and $c \in C$ is a measure $P_c[D]$. The set C contains all the measures for those a conditional measure $P_c[Y/x]$ exists $\forall x \in X$.

Hereinafter we shall use square parentheses to indicate that the measure is defined on some σ-algebra of subsets of the set held, i. e. $P_c[D] : \mathrm{A} \to [0,1]$, where $\mathrm{A} \subseteq 2^D$ – a σ-algebra.

For building a deciding function there is a random independent sample $v_c = \left\{ \left(x^i, y^i\right) \in D \,\middle|\, i = \overline{1, N} \right\}$ from distribution $P_c[D]$ used.

An empirical risk will be sample risk estimation: $\tilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f\left(x^i\right)\right)$.

For the all practically used classification algorithms an empirical risk appears biased risk estimation, being always lowered, as far as the algorithms minimize an empirical risk. So, estimating this bias is actual.

Let $$F(c, Q) = ER\left(c, f_{Q,v}\right), \quad \tilde{F}(c, Q) = E\tilde{R}\left(c, f_{Q,v}\right).$$

Here $Q : \{v\} \to \{f\}$ is an algorithm building deciding functions, and $f_{Q,v}$ – a deciding function built on the sample v by the algorithm Q.

An expectation is calculated over the all samples of volume N.

Introduce an extreme bias function:

$$S_Q\left(\tilde{F}_0\right) = \hat{F}_Q\left(\tilde{F}_0\right) - \tilde{F}_0, \qquad (1)$$

where $\hat{F}_Q\left(\tilde{F}_0\right) = \sup_{c:\tilde{F}(c,Q)=\tilde{F}_0} F(c, Q)$.

We use a supremum because a distribution c is unknown and we assume the "worst" case.

Multinomial Case

In [Nedel'ko, 2003] there is reported the dependency $S_Q\left(\tilde{F}_0\right)$ for the multinomial case when X is discrete, i. e. $X = \{1,\ldots,n\}$, and Q minimizes an empirical risk in each $x \in X$.

For the further comparison let's remember a dependency $S_Q\left(\tilde{F}_0\right)$ in asymptotic case: $\frac{N}{n} = M = \mathrm{const},\ N \to \infty$, $n \to \infty$. Though this is an asymptotic case, the results are applicable to real tasks because the asymptotic bias dependency is close to one for finite samples.

This asymptotic approximation is wholly acceptable already by $n = 10$, herewith it has only one input parameter M.

First, consider "deterministic" case when $\tilde{F}_0 = 0$. In this case $S_Q(0) = \begin{cases} e^{-M/2}, & M \leq 1 \\ \frac{1}{2Me}, & M \geq 1 \end{cases}$.

In general case of $\tilde{F}_0 > 0$ there is no simple analytical formula for $S_Q\left(\tilde{F}_0\right)$ and this dependence is given by plot.

Estimates by Vapnik and Chervonenkis

Now we can calculate an accuracy of Vapnik–Chervonenkis evaluations for the considered case of discrete X, as far as we know an exact dependency of average risk on the empirical risk for the "worst" probabilistic measure.

For $S(\widetilde{F}_0)$ in [Vapnik, Chervonenkis, 1974] there is reported an estimate $S'_V(\widetilde{F}_0) = \tau$, as well as an improved estimate: $S'_V(\widetilde{F}_0) = \tau^2\left(1+\sqrt{1+\frac{2\widetilde{F}_0}{\tau^2}}\right)$, where τ asymptotically tends to $\sqrt{\frac{\ln 2}{2M'}}$, $M' = M/(1-e^{-M})$.

By substitution $\widetilde{F}_0 = 0$ there is resulted $S'_V(0) = \frac{\ln 2}{M'}$.

Let's perform a simple inference of the last formula.

Consider a difference between risk and empirical risk:

$$P\left(\left|R-\widetilde{R}\right| > \varepsilon\right) = P\left(\widetilde{R}=0/R=\varepsilon\right) = (1-\varepsilon)^N .$$

Since the algorithm minimizes an empirical risk, it maximizes the distance between risks:

$$P\left(\sup_{f\in\Phi}\left|R-\widetilde{R}\right| > \varepsilon\right) < |\Phi|(1-\varepsilon)^N ,$$

where Φ is a set of all decision functions. This step implies a replacement of a probability of a sum by the sum of probabilities that makes the main contribution to VC-estimates inaccuracy. Assume right term to be equal to 1 (all probabilistic levels are asymptotically equivalent) and take logarithms:

$$\ln|\Phi| + N\ln(1-\varepsilon) = \ln 1 .$$

Since $|\Phi| = 2^{n\left(1-e^{-M}\right)}$ and $\ln(1-\varepsilon) \approx -\varepsilon$ obtain:

$$S'_V(0) = \varepsilon = \frac{\ln 2}{M'} .$$

Factor $1-e^{-M}$ is a non-zero numbers probability from Poisson distribution and it appears because only "non-empty" values x contribute to capacity.

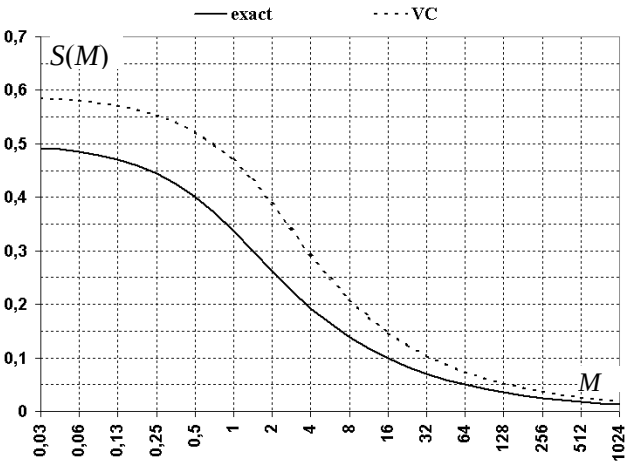

Fig. 1. Risk bias and VC–estimation. Multinomial case, $ER = 0{,}5$.

A rate: $\frac{S'_V(0)}{S_Q(0)} = \frac{2Me\ln 2}{M'} \xrightarrow[M\to\infty]{} 2e\ln 2 \approx 3{,}77$

shows how far VC-estimates are off.

It is known that VC-estimates may be improved by using entropy as a complexity measure. Then the estimate inaccuracy will be:

$$\frac{S''_V(0)}{S_Q(0)} = 2(e-1)\ln 2 \approx 2{,}38 .$$

But in real tasks, entropy can't be evaluated and the last improvement has no use in practice.

On figure 1 there are drawn the dependency $S(M) = \max_{\widetilde{F}_0} S(\widetilde{F}_0)_M$ and its estimation

$$S_V(M) = \max_{\widetilde{F}_0} S_V(\widetilde{F}_0)_M = \sqrt{\frac{\ln 2}{2M'}} .$$

Plots demonstrate significant greatness of the last. Note that the accuracy of Vapnik–Chervonenkis estimation falls since $\widetilde{F}_0$ decreases.

By $M \le 1$ the "worst" distribution (that provides maximal bias) is uniform on X and the results obtained is consistent with results for multinomial case reported in [Raudys, 2001]. By $M > 1$ and restricted $\widetilde{F}_0$ the "worst" distribution is not uniform on X.

Nearest Neighbors Method

This method assigns to each x a class that the most of nearest sample neighbours belongs to.

The number of neighbour objects taken into account is a parameter m that affects a statistical robustness.

Assume a measure on D to be uniform. Then misclassification probability for any decision function is 0,5 and empirical risk is: $\widetilde{F}(m) = \frac{1}{2} - \mathrm{C}_{m-1}^{\left[\frac{m-1}{2}\right]} \frac{1}{2^m}$.

Here square parentheses denote an integer part of a value.

Figure 2 shows $S(M')$ for multinomial case (solid line) and $S(m) = 0{,}5 - \widetilde{F}(m)$ for nearest neighbours classifier, where $m = M'$.

Note that though there is no capacity concept defined for nearest neighbours method the number of neighbours m plays a role of M'.

So the case $m = 1$ corresponds to unbounded capacity (when a sample can be split via decision functions by all the ways). If capacity is unbounded, we can say nothing about expected risk using empirical risk only. But it does not mean that unbounded capacity methods can not be used, it means that they must use other risk estimators.

The fact that a risk bias for multinomial case is close to bias for nearest neighbours classifier is not accidental, because analytic expression for the first one appears to be some kind of averaging the bias for the second case.

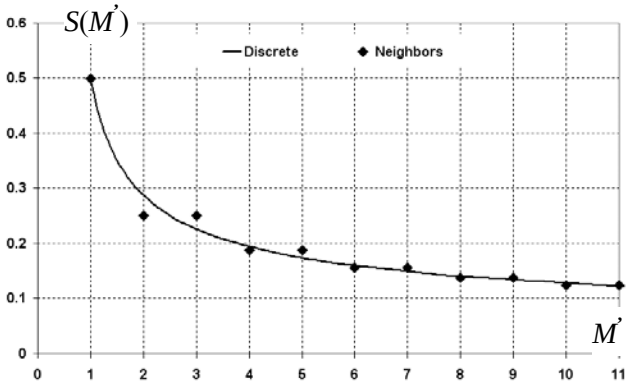

Fig. 2. Risk biases for multinomial and nearest neighbours classifiers.

Linear Decision Functions

Let us compare risk bias values for discrete case with bias for linear decision functions.

For simplifying, there was considered uniform distribution on features for both classes. For such c misclassification probability equals to 0.5 for every decision function, but empirical risk appears to be much lower.

To find a dependence $S(M)$ for linear deciding functions in $X = [0,1]^d$ a statistical modelling was used. By the modelling there was for each combination of parameters a hundred of samples drawn from uniform distribution on D, for each sample the best linear classifier built by exhaustive search. Note that the uniform distribution on D provides maximum of empirical risk bias since we put no restrictions on $\widetilde{F}_0$.

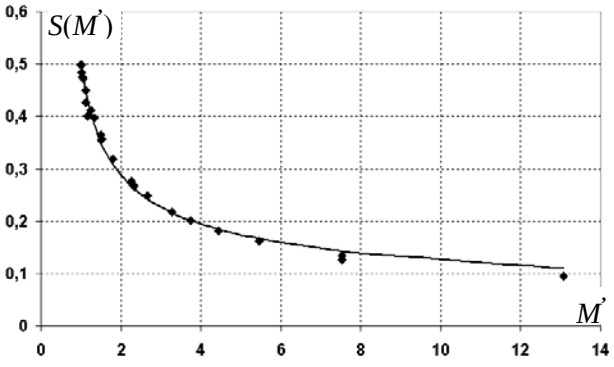

Fig. 3. Risk biases for multinomial and linear classifiers.

A table 1 shows the result of modelling. Here d – features space X dimensionality, N – sample size, $M' = \frac{N}{\log_2 \mathbf{C}}$ – sample size divided by VC-capacity of linear functions class ($\mathbf{C} = 2\sum_{m=0}^{d} \mathrm{C}_{N-1}^{m}$ is a total number of possible decision assignments to sample points by using linear decision functions), S – risk bias.

The same results are shown (by markers) on fig. 3 in comparison with $S(M')$ for discrete case (solid line).

Obtained results show that bias dependence on M' for linear functions is close to dependence for discrete (multinomial) case.

If an algorithm does not perform exhaustive search then a risk bias appears to be lower. This fact is illustrated in table 1 by value S_F that is a risk bias for the Fisher's discriminator.

Decision Tree Classifier

The goal now is to evaluate a risk bias for decision functions in form of binary decision trees [Lbov, Startseva, 1999].

Tab. 1. Risk bias for linear decision functions

d	N	M'	S	S_F	d	N	M'	S
1	3	1.16	0.4	0.4	1	10	2.31	0.27
1	20	3.75	0.2	0.2	1	50	7.53	0.13
1	100	13.1	0.1	0.1	2	4	1.05	0.47
2	10	1.53	0.36	0.27	2	20	2.33	0.27
2	50	4.44	0.18	0.13	2	100	7.53	0.13
3	5	1.02	0.48	0.35	3	10	1.25	0.41
3	20	1.79	0.32	0.2	3	50	3.28	0.22
3	100	5.46	0.16	0.09	4	10	1.11	0.45
4	20	1.5	0.36	0.19	4	50	2.66	0.25
5	10	1.04	0.48	0.27	5	50	2.27	0.28

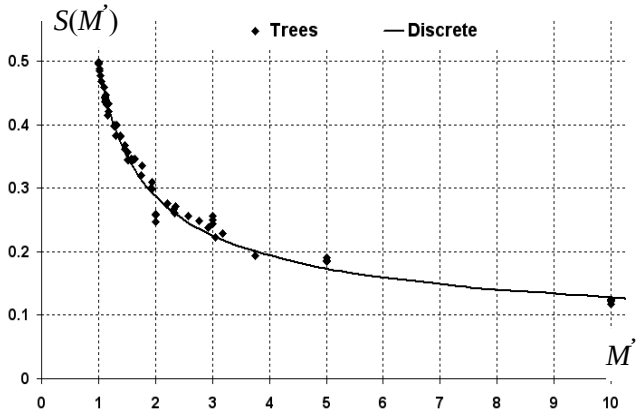

Fig. 4. Risk biases for multinomial and tree classifiers.

Decision tree is a binary tree with terminal nodes marked by goal class (certain value *y*) and non-terminal nodes marked by predicates in form: $X_j < \alpha$, where α is a value. Two arcs starting from each non-terminal node correspond to true and false predicate values.

Each decision tree forms certain sequential partitioning in X.

There was the exhaustive search algorithm implemented. The search is performed over the all decision trees with L terminal nodes and the best tree minimizing an empirical risk is founded.

While searching, the algorithm counts $\mathbf{C}$ – the number of different assignments *y* to sample objects.

Since $\mathbf{C}$ essentially differs on different samples one need to evaluate entropy $H = \mathrm{E} \log_2 \mathbf{C}$.

Then $M' = \frac{N}{H}$.

Tab. 2. Risk bias for tree decision functions

d	N	L	M'	S	d	N	L	M'	S
1	2	1	2	0.26	1	2	2	1	0.5
1	5	2	1.51	0.36	1	5	3	1.12	0.44
1	10	2	2.31	0.27	1	10	3	1.53	0.34
1	20	2	3.76	0.19	1	20	3	2.33	0.26
1	20	5	1.50	0.34	2	5	2	1.26	0.40
2	5	3	1.02	0.49	2	10	2	1.92	0.30
2	10	3	1.28	0.40	2	20	2	3.19	0.23
2	20	3	1.94	0.31	2	20	4	1.46	0.37
3	5	2	1.17	0.42	3	20	2	2.92	0.24
3	20	3	1.77	0.34	3	20	5	1.12	0.45
4	20	2	2.76	0.25	5	10	2	1.57	0.35

Table 2 shows statistical robustness of decision trees by different parameters while uniform distribution on D assumed. The same result is shown on figure 4 in comparison with multinomial case.

One can see again that risk bias is caused and determined by M' (sample size per complexity) rather than any other factor.

Let's compare complexities (capacities) of decision trees and linear classifier.

Table 3 shows linear classifier dimensionality d' that provides the same entropy (average number of different assignments *y* to sample objects) like decision trees with L terminal nodes in d-dimensional space.

Though decision trees seem to be simple, they have essential capacity. For example if $L = d$ decision trees capacity exceeds capacity of linear classifier.

But, the most of algorithms do not perform exhaustive search in whole class of decisions and their capacities are expected to be lower.

Note that if an algorithm implements good heuristic search and always founds the best decision function, then its capacity will be nevertheless equal to the capacity of exhaustive search algorithm. So, there is no use to count a number of decisions being really tested by an algorithm, because this number is irrelevant to actual capacity.

Hence, calculation of effective capacity requires different approach. Effective algorithm capacity may be estimated by the following way.

First one need to perform statistical modelling using uniform distribution on D. In this case misclassification probability (risk) equals to 0,5 for any decision function. Expectation of empirical risk is estimated by modelling, so risk bias is estimated too.

Then via comparing the bias obtained by modelling with the bias for exhaustive search algorithm, the effective capacity of the algorithm under investigation is easily revealed.

Tab. 3. Correspondent dimensionality for tree and linear decision functions. Non-integer values of d' appears because of interpolation performed.

d	*N*	*L*	*d'*	*d*	*N*	*L*	*d'*
1	5	2	1	2	5	2	1.56
2	10	2	1.4	2	20	2	1.3
3	2	2	1	3	5	2	1.83
3	10	2	1.64	3	20	2	1.47
4	5	2	2.09	4	20	2	1.59
5	10	2	1.93	10	10	2	2.45
1	5	3	2	2	5	3	2.95
2	10	3	2.86	2	20	3	2.66
3	5	3	3.76	3	10	3	3.48
3	20	3	3.07	4	5	3	3.99
4	10	3	3.94	2	5	4	3.99
2	20	4	4.26	3	5	4	4
3	10	4	5.82	3	20	4	5.1
4	10	4	6.77	1	10	5	4
2	10	5	6.45	3	15	5	7.77

Conclusion

Risk estimates by Vapnik and Chervonenkis are known to be excessively pessimistic. But the approach based on complexity measure is very attractive because of universality. The work presented shows that the reason for such pessimistic estimates is an inaccurate inference technique, but not the worst case orientation. So, it is possible to obtain estimates assuming the “worst” distribution and the ‘worst’ sample but these estimates will be appropriate in practice.

For the multinomial case (a discrete feature) there was found how far Vapnik–Chervonenkis risk estimations are off. For continuous features the dependence of risk bias on complexity in considered cases is close to multinomial one that ensures a possibility to apply obtained scaling of VC-estimates to real tasks, e.g. linear decision functions and decision trees. The results obtained for multinomial case may be propagated on continuous one by using VC-capacity of decision function class instead of n.

Comparison of linear classifier and decision trees capacities is also performed.

There was also described a method for estimation an effective capacity of an algorithm that does not perform exhaustive search in the class of decision functions.

Bibliography

[Vapnik, Chervonenkis, 1974] Vapnik V.N., Chervonenkis A. Ja. Theory of pattern recognition. Moscow “Nauka”, 1974. 415p. (in Russian).

[Raudys, 2001] Raudys S., Statistical and neural classifiers, Springer, 2001.

[Lbov, Startseva, 1999] Lbov G.S., Startseva N.G. Logical deciding functions and questions of statistical stability of decisions. Novosibirsk: Institute of mathematics, 1999. 211 p. (in Russian).

[Nedel’ko, 2003] Nedel’ko V.M. Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. 2003. pp. 182–187.

Author's Information

Victor Mikhailovich Nedel’ko – Institute of Mathematics SB RAS, Laboratory of Data Analysis, 660090, pr. Koptyuga, 4, Novosibirsk, Russia, e-mail: nedelko@math.nsc.ru