Misclassification Probability Estimations for Linear Decision Functions

Victor Mikhailovich Nedel'ko

Institute of Mathematics SB RAS, Laboratory of Data Analysis

660090 Novosibirsk, Russia

nedelko@math.nsc.ru

<http://math.nsc.ru/LBRT/i1/nedelko/index.html>

Abstract. The work is devoted to a problem of statistical robustness of deciding functions, or risk estimation. By risk we mean some measure of decision function prediction quality, for example, an error probability. For the case of discrete “independent” variable the dependence of average risk on empirical risk for the “worst” distribution (“strategies of nature”) is obtained. The result gives exact value of empirical risk bias that allows evaluating an accuracy of Vapnik–Chervonenkis risk estimations. To find a distribution providing maximum of empirical risk bias one need to solve an optimization problem on function space. The problem being very complicate in general case appears to be solvable when the “independent” feature is a space of isolated points. The space has low practical use but it allows scaling well-known estimations by Vapnik and Chervonenkis. Such scaling appears to be available for linear decision functions.

1 Introduction

There is well known fact that decision function quality being evaluated by the training sample appears much better than its real quality. To get true risk estimation in data mining one uses a testing sample or moving test. But these methods have some disadvantages. 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 dispersion.

So one need a method that allows estimating a risk by training sample directly, i. e. by an empirical risk. This requires estimating first an empirical risk bias.

The problem was solved by Vapnik and Chervonenkis [1]. They introduced a concept of capacity (growth function) of a decision rules set. This approach is quite powerful, but provides pessimistic decision quality estimations. Obtained bias estimation is very rough, because of performed by authors replacement of a probability of a sum of compatible events by the sum of its probabilities.

The goal of this paper is to evaluate an accuracy of these estimations. We shall find out how far they are off by considering a case of discrete feature that allows obtaining an exact value of empirical risk bias.

2 Problem Definition

Let X – an “independent” variable values space, Y – a goal space of forecasting values, and C – a set of probabilistic measures on $D = X \times Y$. A measure $c \in C$ will be $P_c[D]$. The set C contains all the measures for those a conditional measure $P_c[Y/x]$ exists $\forall x \in X$.

Hereinafter square parentheses will hold a set on which σ -algebra of subsets the measure assigned, but parentheses — the measure of the set (probability of event).

A deciding function is a correspondence $f : X \rightarrow Y$.

For the determination of deciding functions quality we need to assign a function of losses: $L : Y^2 \rightarrow [0, \infty)$.

By a risk we shall understand an average loss:

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

Another form is:

$$R(c, f) = \int R_x(c, f) dP_c[X], \text{ where } R_x(c, f) = \int L(y, f(x)) dP_c[Y/x].$$

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

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

For the all practically used algorithms building deciding functions an empirical risk appears a biased risk estimation, being always lowered, as far as algorithms minimize an empirical risk.

So, estimating this bias is actual.

Enter indications:

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

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

Expectation is calculated over the all samples of volume N .

Introduce an extreme bias function:

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

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

One of the results of this work consists in finding the dependency $S_Q(\tilde{F}_0)$ for the multinomial case when X is discrete and Q minimizes an empirical risk in each $x \in X$.

3 Discrete Case

Let X be discrete, i. e. $X = \{1, \dots, n\}$.

Then $R(c, f) = \sum_{x=1}^n p_x R_x(\xi_x, f(x))$, where $R_x(\xi_x, v_x) = \int L(y, f_{v_x}(x)) dP_c[Y/x] - a$ conditional risk in the point x , $p_x = P_c(x)$, ξ_x – a short indication of conditional measure $P_c[Y/x]$, $v_x = \left\{ (x^i, y^i) \in v \mid x^i = x \right\}$, $f_{v_x}(x)$ – a decision built on sub-sample v_x .

Let the deciding function minimize an empirical risk: $f_v^*(x) = \arg \min_{y \in Y} \tilde{L}(y, v_x)$,

$$\tilde{L}(y, v_x) = \frac{1}{|v_x|} \sum_{v_x} L(y, y^i).$$

Determine values which average risks depend on:

$$F(c, Q) = ER(c, f_v^*) = \sum_{x=1}^n F_x(p_x, \xi_x), \text{ where } F_x(p_x, \xi_x) = p_x ER_x(\xi_x, f_{v_x}^*(x)).$$

Similarly:

$$\tilde{F}(c, Q) = \frac{1}{N} E\tilde{L}(v, f_v^*) = \sum_{x=1}^n \tilde{F}_x(p_x, \xi_x), \text{ where } \tilde{F}_x(p_x, \xi_x) = E\tilde{L}_x(v_x, f_{v_x}^*(x)).$$

Introduce the function

$$\hat{F}_x(p_x, \tilde{F}_x^0) = \sup_{\xi_x: \tilde{F}_x(p_x, \xi_x) = \tilde{F}_x^0} F_x(p_x, \xi_x).$$

Now

$$\hat{F}_Q(\tilde{F}_0) = \max_{x=1}^n \hat{F}_x(p_x, \tilde{F}_x^0), \quad (2)$$

where maximum is taken over the all p_x and $\tilde{F}_x^0$, $x = \overline{1, n}$ with restrictions: $p_x \geq 0$,

$$\tilde{F}_x^0 \geq 0, \quad \sum_{x=1}^n p_x = 1, \quad \sum_{x=1}^n \tilde{F}_x^0 = \tilde{R}_0.$$

Initial extreme problem is rather simplified, being split into two parts: finding $\hat{F}_x(p_x, \tilde{F}_x^0)$ and finding the maximum of function on the simple area of Euclid space.

Function $\hat{F}_x(p_x, \tilde{F}_x^0)$ may be easily approximated by calculation on a grid.

However it is impossible to solve a problem (2) directly by digital methods, because the dimensionality of space on that the maximization is performed is $2n$ where n may be great (twenty and the more).

4 Deciding the Extreme Problem

Let's reformulate a problem (2) in abstract indications:

$$\sum_{x=1}^n \Phi(z^x) \rightarrow \max_{z^x}, \quad (3)$$

$$z^x = (z_1^x, \dots, z_m^x), \quad z_j^x \geq 0, \quad \sum_{x=1}^n z_j^x = 1, \quad j = \overline{1, m}.$$

In (2) Φ corresponds to $\hat{F}_x$, $m = 2$, $z_1^x = p_x$, $z_2^x = \frac{\tilde{F}_x^0}{\tilde{F}_0}$.

Suppose a space of values of vector z^x to be made discrete: $z^x \in \{t^1, \dots, t^l\}$.

Then problem (3) now may be put in the equivalent form:

$$\sum_{i=1}^l \Phi(t^i) \kappa^i \rightarrow \max_{\kappa^i}, \quad (4)$$

$$\kappa^i \geq 0, \quad \sum_{i=1}^l t_j^i \kappa^i = 1, \quad j = \overline{1, m}, \quad \sum_{i=1}^l \kappa^i = 1, \quad \text{where } \kappa^i \in \left\{ \frac{\nu}{l} \mid \nu = \overline{0, l} \right\}.$$

Solve a problem without the last restriction on κ^i .

This is a linear programming problem with a decision in a vertex of values area.

This means that only $m+1$ of κ^i -s may be nonzero.

Now it is easy to show that the effect of discreteness κ^i restriction has an influence like $\frac{m}{l}$ and one may neglect it.

As far as the conclusion on the number of nonzero κ^i -s does not depend on the step of sampling a space of vector z^x values, the result may be propagated to the initial problem.

Theorem 1. Decision of problem (3) includes not more than $m+1$ different vectors.

Thereby the dimensionality of space on which maximizing is performed decreases to $m(m+1)$.

Applying to the problem (2) this comprises 6, and problem may be easily solved numerically.

5 Results

Offered method allows finding a dependency $S_Q(\tilde{F}_0)$ by any parameters n and N .

However there is the most illustrative and suitable for the comparison with other approaches the asymptotic case: $\frac{N}{n} = M = \text{const}$, $N \rightarrow \infty$, $n \rightarrow \infty$.

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

For illustrations consider a problem of categorization (two classes).

The loss function is: $L(y, y') = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$.

Now we are to calculate an accuracy of Vapnik–Chervonenkis evaluations for the considered case of discrete X , as far as we have found an exact dependency of average risk on the empirical risk for the “worst” strategy of nature.

Fig. 1. The exact dependency of supreme risk bias on empirical risk and its estimations are shown by different M . Solid line shows difference between expected misclassification error and empirical risk, the difference maximized over the all distributions providing expected empirical risk $\tilde{F}_0$.

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

On figure 1 for the case of two classes by different M there are drawn the dependency $S(\tilde{F}_0)$ and its estimation $S_V(\tilde{F}_0)$. Plots demonstrate significant greatness of the last. Note that the accuracy of Vapnik–Chervonenkis estimation falls since the sample size (parameter M) decreases.

By $M \leq 1$ the “worst” distribution (that provides maximal bias) is uniform on X and the results obtained is consistent with results for multinomial case reported in [2]

(table 3.7). By $M > 1$ and restricted $\tilde{F}_0$ the “worst” distribution is not uniform on X and this fact is new.

The third line on graphs (straight) presents a variant of dependency obtained by Byesian approach, based on assuming the uniform distribution on strategies of nature [3,5].

In work [5] there is reported the result: $S_B(\tilde{F}_0) = \frac{1 - 2\tilde{F}_0}{M + 2}$.

It may be a surprised fact that $S(\tilde{F}_0)$ on the significant interval appears to be greatly less than $S_B(\tilde{F}_0)$, though Byesian approach implies an averaging by the all nature strategies of nature, but our approach — the choice of the worst c .

6 Linear Decision Functions

Let us compare risk bias values obtained 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.

We can evaluate an accuracy of Vapnik-Chervonenkis risk estimations for the case of discrete X , as far as we know an exact dependency of average risk on the empirical risk for the “worst” strategy of nature (distribution).

On fig. 2 for the case of two classes and discrete X there are drawn the dependency $S(M) = 0.5 - \tilde{F}(c_U)$ and its estimation $S_V(M) = \sqrt{\frac{\ln 2}{2M'}}$ by Vapnik and Chervonenkis. Here $\tilde{F}(c_U)$ – expected empirical risk by uniform distribution; $\frac{N}{n} = M = \text{const}$, $N \rightarrow \infty$, $n \rightarrow \infty$; $M' = M / (1 - e^{-M})$ is sample size divided by VC-capacity of decision functions class in discrete case. Note that the uniform distribution on D provides maximum of empirical risk bias since we put no restrictions on $\tilde{F}_0$.

Fig. 2. Risk bias and VC-estimation. Discrete case, ER = 0.5.

Fig. 3. Risk biases for multinomial and linear decision functions.

Graphics demonstrate significant greatness of Vapnik-Chervonenkis estimation.

To find a dependence $S(M)$ for linear deciding functions in $X = [0,1]^d$ a statistical modeling was used. By the modeling 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.

A table shows the result of modeling. Here d – features space X dimensionality, N – sample size,

$$M' = \frac{N}{\log_2 C} \text{ – sample size divided by VC-capacity}$$

of linear functions class (C 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.

7 Conclusion

For the considered case of a discrete feature the exact maximum of empirical risk bias have been obtained. The comparison conducted shows that risk estimations by Vapnik and Chervonenkis may increase an expected risk up to several times from its true maximum. This means that these estimations may be essentially improved.

d	N	M'	S
1	3	1.16	0.4
1	5	1.5	0.36
1	10	2.31	0.27
1	20	3.75	0.20
1	50	7.53	0.13
1	100	13.08	0.095
2	4	1.05	0.47
2	10	1.53	0.357
2	20	2.33	0.27
2	50	4.44	0.18
2	100	7.53	0.13
3	5	1.019	0.48
3	10	1.247	0.41
3	20	1.79	0.32
3	50	3.28	0.22
3	100	5.46	0.162
4	6	1.008	0.498
4	10	1.111	0.45
4	20	1.5	0.355
4	50	2.66	0.249
5	7	1.0033	0.499
5	10	1.044	0.476
5	20	1.33	0.398
5	50	2.27	0.275

Obtained improvement is applicable also for continuous space, e.g. linear decision functions.

Practical use of the result consists in that one can apply obtained scaling of VC-estimations to real tasks. The results obtained for multinomial case may be propagated on continuous one by using VC-capacity of decision function class instead of n .

The work is supported by RFBR, grant 04-01-00858-a.

References

1. Vapnik V.N., Chervonenkis A. Ja. Theory of pattern recognition. Moscow "Nauka", 1974. 415p. (in Russian).
2. Raudys S., Statistical and neural classifiers, Springer, 2001.
3. 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).
4. Nedelko V.M. An Asymptotic Estimate of the Quality of a Decision Function Based on Empirical Risk for the Case of a Discrete Variable. // Pattern Recognition and Image Analysis, Vol. 11, No. 1, 2001, pp. 69-72.
5. Berikov V.B. On stability of recognition algorithms in discrete statement. // Artificial Intelligence. Ukraine, 2000, N. 2. P. 5–8. (in Russian).