INDUCTIVE METHOD PERMITTING TO CHOOSE MODEL WITH LEAST ERROR AND LEAST BIAS ALLOWING THE SOLVE INTERPOLATION TASKS OF ARTIFICIAL INTELLECT

A.G.Ivakhnenko[1], E.A.Savchenko[1], G. A.Ivakhnenko [1],

A.B.Nadiradze[2], A.O.Rogov[2]

The set of object subjecting to recognition or dependence detection can be closed i.e. limited. In this case it is reasonable least error to use external criterion of in order to choose optimal model among given collection of model-candidates. In this case if data is accurate enough, model selection is ambiguous. The interval of model ambiguous is formed due to the fact that model complexity is changed discretely as a dependence of argument amount. Then it is recommended to use search over criterion of bias for model additional determination. It is suggested and used a new cross criterion of model bias, to calculate which is easer than to calculate criterion of model bias with the help of retrieval separation onto two statistically identical parts.

INTRODUCTION

Interpolation tasks of artificial intelligence are solved or deductive, through purpose of complexity of structure of model by the man - author of modeling [1] or inductive way, with the help search of set of the models - candidates by external criteria. In the second case it is possible to name algorithm as self-organizing of model [2,3].

The systems of interpolation type, in which model or discriminate of function are chosen from set of the models - candidates only by one criterion of the least mistake, are suitable only for the closed set of entrance objects, for which it is possible to specify the standards. Such systems concern to group of the systems, which have received the name of systems of search of the data (*data mining*).

If the set of the standards continuously extends, for example, at recognition of hand-written marks, it is necessary to pass to models having property of generalization, i.e. exact enough on samples, which can be received on the same object, but with which the author of modeling has no. Such models, similarly to the laws of physics, should be unbiased, i.e. important for a choice of optimum model there is a criterion of bias. The task of a choice of model is solved successfully, if the model exact enough with small enough bias is found. The choice of such model concerns to area of extraction of new knowledge (*knowledge extraction*), since thus there is new knowledge of the object, which has been not specified in initial sample of the data.

---

[1]International Scientific and Educational Center of Information Technologies and Systems of National Academy of Science of Ukraine, 03187, Kiev, Pr. Glushkova, , 40, (044)266 30 28, koleso@i.kiev.ua

[2] Moscow aviation institute (technical university), Moscow, 125993, Volokolamskoe shosse, 4, MAI, (095)158-4674, Andrey.Nadiradze@mtu-net.ru

HOW TO CHOOSE ONE OPTIMAL MODEL ACCORDING TO COMBINE CRITERION OR TWO SERIAL OPTIMIZATIONS WITH ADDITIONAL DETERMINATION OF THE MODEL

The absence of bias requires, that two models received on two identical samples of the data (on average and on dispersion variable), should be identical on all parameters. This requirement can be expressed mathematically by different ways. For example, in space of factors polynomial of models the criterion of bias will be such:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_Mx_M.$$

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Mx_M.$$

$$a_0 = b_0,\ a_1 = b_1,\ a_2 = b_2,\ \ldots,\ a_M = b_M.$$

The criterion of bias designed in space of mistakes is more convenient:

$$ER = RR_{A/A+B} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \rightarrow \min$$

$$BS = |RR_{A/A+B} - RR_{B/A+B}| \rightarrow \min$$

Combine criterion, taking into account either error or model bias can be written as:

$$\rho = \sqrt{\lambda \cdot ER^2 + (1-\lambda) \cdot BS^2} \rightarrow \min,$$

where λ – weight coefficient, setting by author of simulation. The two models, described before, correspond to boundary cases $\lambda=0$; $\lambda=1$. If one likes not to give weight coefficient, it is possible to use serial optimization over one criterion and then over other one (It is correct as far as model choice over error and over bias is orthogonal, i.e. criterions are independent). In this case the idea of model self-organization is not spoiled. For example, it is possible to find small set of models, which are the best according to criterion of error, and then to select one mode, optimal according to criterion of bias. It is good that when one search modes, the interval of ambiguous of model choice **LC-RC** is formed in search performance due to the fact that complicity of model, the representative points of which are in this interval. In this case two serial optimizations can be called as additional determination of the model. The bias of the model is used 3 times in the algorithm: 1) in the indirect form in combine algorithm (when coefficient estimation is determine for retrieval ***A*** and when model optimal structure is chosen for retrieval ***B***; 2) in direct form when the most undisplaced mode is determined additionally among several models; 3) when one model with least bias is chosen.

If the sample of the data contains continuous meanings variable, the division her on two samples with equal statistical properties (average and дисперсией) is complicated. In this case it is more convenient to apply cross criterion of a mistake (Cross-Validation) and cross criterion of bias (Cross-Bias Criteria). The account of cross criterion of bias serves as though continuation of account of cross criterion of a mistake of model. It can be described so.

For account of cross criterion of bias we exclude lines consistently on one. On the excluded lines we determine a square of a mistake estimated on bias of model. At number of lines of sample equal N, we shall receive as much of squares of a mistake of one line: $er_{12}, er_{22}, \ldots, er_{N2}$. Further we consider meaning of squares of a mistake for all other lines, except for one, excluded. Let's receive still N of meanings of squares of an average mistake:

$$ER_1^2, ER_2^2, ..., ER_N^2,\ E\overline{R} = \frac{1}{N}\sum_{i=1}^{N} er_i\ ,\ i = 1, 2, \ldots, \text{N}.$$

If the bias of model is equal to zero, the squares of a mistake on one line should be equal to average meaning of squares of mistakes on all other lines. The difference of the sums of squares of mistakes enables to estimate bias of model under the formula:

$$BS^2 = \frac{1}{N}\sum_{i=1}^{N}(er_i^2 - E\overline{R}^2) \rightarrow \min,\ i = 1, 2, \ldots, \text{N}.$$

The cross criterion of bias can be applied both for an estimation of bias of models, and for automatic clusterization of sample of the data on кластеры, necessary, for example, for multialternative recognition of images and classification.

HOW TO CHOOSE OPTIMAL MODEL IN THE ALGORITHM PERMITTING TO DETECT SPUTTERING COEFFICIENT DEPENDENCE ON MATERIAL'S PHYSICAL PROPERTIES

Physical idea of sputtering coefficient dependence on prescribed arguments is represented in [4]. Initial data retrieval, normalized over the largest value of every variable is represented in the table 1. Here x_1 – sputtering coefficient (mg/C) by Xe^+, 300 eV; x_2 – mass density (g/cm^3); x_3 – molecular weight; x_4 – temperature of sublimation; x_5 – heat-capacity (J/mole/degree); x_6 – energy of bonds (eV). Output variable is x_1. The result of finding will be all regularities with sufficiently small displacement.

In the given example, as well as in previous, four models are received: by two deductive method and two through self-organizing on to combinatory algorithm GMDH with additional determination on bias. The results are submitted in a figure and in the tables 2 and 3.

Table 1. Initial data.

	x1	x2	x3	x4	x5	x6
C	0,0068	2,3	12,0	4473	8,5	7,4
Be	0,0228	1,8	9,0	2744	16,4	3,4
Si	0,1172	2,3	28,1	3573	19,7	3,9
Ti	0,1175	4,5	47,8	3560	25,0	4,3
V	0,1409	5,9	50,9	3665	24,4	3,7
Cr	0,2310	7,2	51,9	2945	23,5	3,6
Nb	0,2322	8,5	92,9	5073	24,4	7,5
Al	0,2407	2,6	26,9	2793	24,3	3,2
Fe	0,3168	7,8	55,8	3145	24,9	4,1
Co	0,3786	8,9	58,9	3230	24,8	4,3
Mo	0,3837	10,2	95,9	4700	23,9	6,9

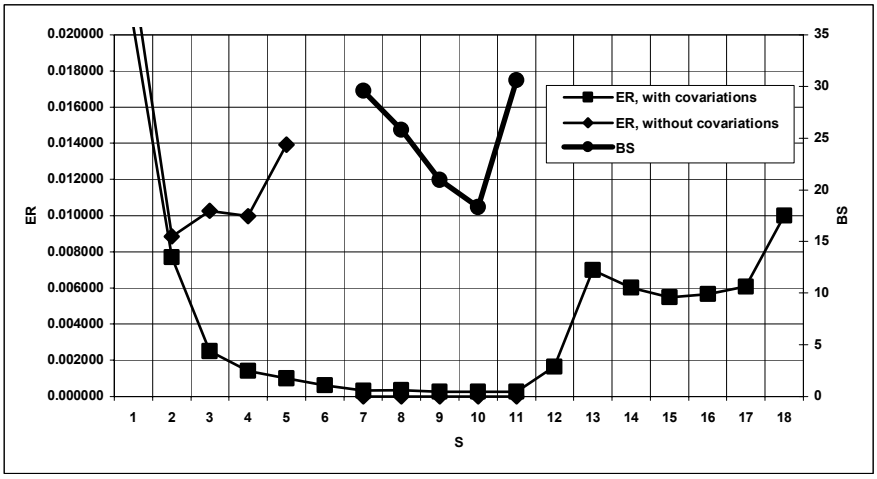
	x1	x2	x3	x4	x5	x6
Ge	0,4355	5,3	72,5	3120	23,2	3,7
Cu	0,6565	8,9	63,5	2816	24,4	3,5
Ta	1,1587	16,6	180,9	5623	25,2	8,7
Pd	1,1706	12,2	106,4	3273	25,8	4,8
Mn	1,2719	7,3	54,9	2353	26,2	3,1
W	1,3408	19,3	183,8	5953	24,2	8,7
Pt	1,6581	21,4	195,1	4100	25,8	5,5
Ag	1,9414	10,5	107,8	2440	25,3	2,7
Au	2,3635	19,3	196,9	3150	25,4	3,9
Max	*2,3635*	*21,4*	*196,9*	*5953*	*26,3*	*8,7*
MCC	1,000	0,810	0,813	0,005	0,464	0,021

Table 2. A choice of model through purpose of limiting meaning of factor of correlation MCC in a task of detection of dependences (without application GMDH)

Arguments	Optimal models	ER	BS
x_i, i = 2,3,4,5,6.	$y = 0,779 – 0,220x_2 + 1,333x_3 –1,719x_4 - 0,255x_5 +0,570x_6$	0,014	0,814;
x_i; x_ix_j; i = 2,3,5; i,j = 2,2; 2,3; 2,4; 2,5; 2,6; 3,3; 3,4; 3,5; 3,6; 5,5.	$y = -0,324 + 9,98x_2 – 12,63x_3 + 0,63x_5 – 14,4x_2^2 + 22,38x_2x_3 – 57,83x_2x_4 + 1,15x_2x_5 + 47,65x_2x_6 – 8,59x_3^2 + 58,43x_3x_4 + 3,38x_3x_5 – 48,71x_3x_6 –0,482x_5^2$	0,0081	0,415

Table 3. A choice of model on combinatory to algorithm GMDH with additional determination on bias in a task of detection of dependences (with application combinatory of algorithm GMDH)

Arguments	Optimal models	ER	BS
x_i, i = 2,3,4,5,6.	$y = 0,779 – 0,220x_2 + 1,3334x_3 –1,719x_4 - 0,255x_5 +0,570x_6$	0,014	0,814;
x_ix_j, i,j = 6; 2,2; 2,4; 2,5; 2,6; 3,4; 3,5; 4,4; 4,5; 4,6.	$y = 1,434 – 5,856x_6 + 0,430x_2^2 –4,435x_2x_4 –1,309x_2x_5 +7,891x_2x_6 –7,189x_3x_4 + 4,781x_3x_5 +0,723x_4^2 – 0,580x_4x_5 + 4,992x_4x_6$	0,00027	0,115.



So, inductive method of self-organizing of model has given more effective results, than deductive method (tab. 2 and 3).

REFERENCES

1. Krug G.M., Krug O.Yu. A mathematical method of classification of ancient ceramics // Works of archaeological institute SA USSR.– Moscow: Nauka, 1965.– c.317 – 323..
2. Ivakhnenko A.G., Ivakhnenko G.A. and Mueller J.-A. Self-Organization of Optimum Physical Clustering of Data Sample for a Weakened Description and Forecasting of Fuzzy Objects. \\ Pattern Recognition and Image Analysis, v.3, N4, 1993, pp. 415-422.
3. A.G. Ivakhnenko, S.A. Petuhova at another. Objective Selection of Optimal Clustering of a Data Sample During Compensation of Non-Robust Random Interference // Journal of Automation and Information Sciences, 26 (3), 1993.
4. Sputtering of solid bodies by ion bombing: Physical sputtering of single-element solid bodies. Editor R.Berish - Mir, 1984, pp. 336.