TRAINING FOR EXTREMUM DETERMINATION OF FUNCTION OF VARIABLES MEASURED IN NAMES SCALE

G.S.Lbov

Laboratory of Pattern Recognition,
Institute of Mathematics,
Siberian Division,
USSR Academy of Sciences,
U.S.S.R.

ABSTRACT

In this paper an algorithm is given for extremum search of an unknown function *F(x)* when the space of variables taking discrete values is not a metric one. Function values are obtained experimentally.

The given algorithm is based on the idea of adaptive random search. This algorithm has been used for the solution of a number of practical problems for choosing an effective subsystem of dependent features in pattern recognition. It is intended for solving problems which do not correspond to know problems of discrete programming.

S 1. Formulating of problem

This paper deals with the following problem: there is a set of n. variables X_1 A'_2, , X_T, ,X_n. Everyone of those takes the finite set of values $X_1 = xtf. .. , x_{t\beta}, . , x_{t\ell_t})$. Here, variables are quantities measured in a names scale (1) In other words, everyone of these value sets is the list of some object names, i.e. the space constructed by these variables is not a metric one.

Let us form the set of *n* elements $X = (x_1 . x_t, ,x_n)$ taking one element from each set. The number of different sets is equal $N = \prod_{i=1}^{n} \ell_i$. For everyone of these element sets one can determine *F(x)*- the quality criterion value of this set. It is required to find such a set $x^*=(x_1^*, x^{**})$ in order for *F(Oc)* to take the extreme value. Thus formulated, the problem can be solved only with a method involving complete sorting, generally speaking. As a rule, it is impossible to carry out the complete sorting because of the great number of variants N and the considerable cost of carrying out the experiment. However, it is possible to sort only a limited number *T* of different variants *(T—N)* . In this case, if there are no assumptions of function class restrictions, one can use any algorithm for function extremum search.

In this paper an algorithm based on the idea of adaptive random search is given. This idea is the following: organization of subsequent tests depends on the results of previous ones. This procedure is the training for extremum determination of the function. For that the number of tests is broken up into a number of groups

$$T = \tau^{(1)} + \tau^{(2)} + \ldots + \tau^{(\varphi)} + \ldots + \tau^{(R)}$$

and the following assumption is used for testing the φ-the group: for any pair of values $\{x_{t\beta}, x_{t\gamma}\}$ of variable X_t the probability $P^{(\varphi)}\{x_{t\beta} = x_t^*\} > P^{(\varphi)}\{x_{t\gamma} = x_t^*\}$ even if one set a? including x is better (according to criterion *'F(x)* than any set x including x_{tv}, . The difference between these probabilities is greater the more is the relative number of tests carried out by the given

moment of search

$$\tau_\varphi = \frac{\sum_{\kappa=1}^{\varphi} z^{(\kappa)}}{T}$$

Tests of the $(\varphi+1)$-th group are carried out in order to include the value $x_{i\beta}$ of variable X proportional to the probability $p^{(\varphi)}\{x_{i\beta} = x_i^*\}$ which is changed as the search is carried out. At the beginning of the search, if there is no a priori information for prefering one value X to another, the probabilities of choice of these values are supposed to be equal to $\frac{1}{\ell_i}$.

The greater the number of tests the more "careful" must be the extremum search and the larger the problem class solving.

This problem does not correspond to known problems of discrete programming.

The problem of choosing the most effective subsystem of dependent features in pattern recognition is an example of such problems for whose solution the algorithm stated below is supposed to be used.

Really, here the set of all possible variants is that of all solutions of including either feature in an effective subsystem. The variable X takes two values: the i-th foature is included in the feature subsystem or not. In this case, the variant number is equal $N=2^n$.

A summary of discrepancies connected with both feature measure and recognition errors can be used, for instance, as the criterion $F(x)$.

The other example of such problems is that of technological process optimization if there are variables measured in the names scale.

The adaptive random search idea has been used for algorithm (3)for the choice of m features from n ($N = C_n^m$). This algorithm has been an effective one for solution of a number of practical problems in pattern recognition, medicine, geology, sociology, psychology, economics.

§2. Algorithm description.

Let us introduce the following definitions:

1. Space of positions.

Let us denote it by $\mathcal{Z}$. The space of positions is constructed by variables $\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_i, \ldots \mathcal{Z}_n$ corresponding to those $X_1, X_2, \ldots, X_i, \ldots, X_n$. Variable $\mathcal{Z}_i$ takes one of the integer values $\{1,2, \ldots, \beta, \ldots, m_i\}$ which we name positions. The number of positions m_i is equal to that of variable values X_i in a given moment of search ($m_i \leq \ell_i$)*.

After carrying out every experiment group, a one-to-one correspondence is established between the positions of variable $\mathcal{Z}_i$ and the values of variable X_i. For that, the best set $x^{(tf)}$ including the value xtf, the best set $x^{(12)}$ including the value x_{12} and so on are fixed. Putting in order sets $x^{(11)}, x^{(12)}, \ldots, x^{(1m)}$ by the criterion we thereby put in order the variable values X_i. The variable value for which the best result by criterion $F(x)$ is obtained is compared to the best position, the value for which the next

As the search is carried out,the number of variable values X_1 used is decreased by excluding variable values of less scope. The choice of number m.will be considered below.

result by quantity is obtained - to second one and so on.

2. Set of transforming vectors. Let us denote it by $\alpha = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n)$. Vector $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{i\beta}, \ldots, \alpha_{i\ell_i})$ establishes a correspondence between values of variable X_t and positions of variable $\mathcal{X}_i$. The component of this vector $\alpha_{i\beta}$ is the value which takes the β-th position in a given moment of search.

At the beginning of the search, if there is no a priori information of preference of one variable value to another the vector is

$$\alpha_i = (x_{i1}, x_{i2}, \ldots, x_{i\beta}, \ldots, x_{i\ell_i}).$$

Since the vectors set is renewed after carrying out each group of experiments, denote it by $\alpha^{(\varphi)}$.

3. Matrix of recommendations $B^{(\varphi)}$. This matrix is a set of recommendations in a space of positions for carrying out a φ-th group experiments.

$$B^{(\varphi)} = \left\{ \mathcal{X}_i^{(\xi)} \right\}$$

for $i = 1,2,\ldots n$; $\xi = 1,2,\ldots z^{(\varphi)}$;

4. Matrix of planning $G^{(\varphi)}$. It is obtained from matrix $B^{(\varphi)}$ using a set of vectors. Matrix of planning is

$$G^{(\varphi)} = \left\{ x_i^{(\varphi)} \right\}$$

for $i = 1,2,\ldots n$; $\xi = 1,2,\ldots z^{(\varphi)}$;

5, Function of "confidence" P. This function is introduced as a quantitative measure of preference of one value to another

$$P = \left\{ P_\beta (b_i, m_i) \right\}$$

for $i = 1,2,\ldots n$; $\beta = 1,2,\ldots \ell_i$; where $P_\beta (b_i, m_i)$ is a probability of variable value X which is occupying the β-th position in given moment of search with the value x_i^*. The choice of parameter b_i will be considered below.

The function of "confidence" is given in a space of positions.

Proceeding from the assumption described in §1 the function $P_\beta (b_i, m_i)$ might have the following properties:

1) at the beginning of search if there is no a priori information of preference of one variable value to another

$$P_\beta (b_i, m_i) = \frac{1}{m_i}$$

2) function $P_\beta (b_i, m_i)$ must decrease steadily when the number of the position increases.

3) by increasing the relative number of experiments carried out, the function of "confidence" must increase at the first positions at the expense of those with greater numbers.

4) $\sum_{\beta=1}^{m_i} P_\beta (b_i, m_i) = 1$

In our case we have used the function $f(x) = \frac{b+1}{(1+bx)^2}$ which is given in paper (4).

For any value of parameter b from interval $(0, \infty)$

$$\int_0^1 f(x)\,dx = 1$$

Function of "confidence" is given thus:

$$P_\beta (b_i, m_i) = \int_{\frac{\beta-1}{m_i}}^{\frac{\beta}{m_i}} \frac{b_i + 1}{(1 + b_i x)^2}\,dx =$$

$$= \frac{b_i + 1}{b_i + \frac{m_i}{\beta}} - \frac{b_i + 1}{b_i + \frac{m_i}{\beta - 1}}. \quad (1)$$

The function (1) has all properties formulated above. Really, if $\delta_i = 0$ then all of m_i positions of variable x_i have equal value for the function of "confidence" - $\frac{1}{m_i}$. If $\delta_i > 0$ then function $P_\beta(\delta_i, m_i)$ is steadily decreasing when the number of position is increasing. If parameter δ_i is increasing as the number of experiments is storing then the values of variable x_i occupying the first positions will be more prefered.

Since the function of "confidence" depends on the group number of experiments carried out, we denote it by $P_\beta^{(\varphi)}(\delta_i, m_i)$.

Now, if the number of experiments corresponding to either value x_i is done in proportion to the function of "confidence" value then the number of experiments will be stored; the more perspective variable values $x_1, x_2, \ldots, x_i, \ldots, x_m$ will be included by further search. The number of experiments $z_{i\beta}^{(\varphi+1)}$ for the variable value keeping the β-th position is done equal to the nearest integer from $z^{(\varphi+1)} P_\beta^{(\varphi)}(\delta_i, m_i)$ i.e.

$$z_{i\beta}^{(\varphi+1)} = \text{окр}\left\{ z^{(\varphi+1)} P_\beta^{(\varphi)}(\delta_i, m_i) \right\}$$

At every stage of search a one-to-one correspondence is established between points of space $\mathcal{X}$ constructed by variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_i, \ldots, \mathcal{X}_n$ and space X constructed by variables $X_1, X_2, \ldots, X_i, \ldots, X_n$. This correspondence is done by means of the vector set $\alpha^{(\varphi)}$. Therefore, if in space $\mathcal{X}$ the set of $z^{(\varphi+1)}$ points is chosen in correspondence with the distribution law given by the function of "confidence" and if condition (2) is satisfied then in space X set of points which also satisfies equation (2) will correspond to that set. Hence, we obtain point sets in space $\mathcal{X}$ (matrixes $B^{(1)}, \ldots, B^{(\varphi)}, \ldots, B^{(R)}$) before carrying out experiments. It is necessary to input in ECM only the following data: the number of variables n, the number of values for all of the variables $l_1, l_2, \ldots, l_i, \ldots, l_n$ and the number of experiments T.

Let us describe the procedure for obtaining matrixes $B^{(1)}, B^{(2)}, \ldots, B^{(\varphi)}, \ldots, B^{(R)}$. At first, it is necessary to determine the function of "confidence" $P_\beta^{(\varphi)}(\delta_i, m_i)$ using formula (1). We give coefficient δ_i and the number of positions m_i the following way: the entropy for variable $\mathcal{X}_i$ is calculated by means of the known correlation

$$H^{(\varphi)}(\delta_i, m_i) = -\sum_{\beta=1}^{m_i} P_\beta^{(\varphi)}(\delta_i, m_i) \cdot \log P_\beta^{(\varphi)}(\delta_i, m_i) \qquad (3)$$

Then the entropy for whole space of positions is determined. This entropy is equal to the sum of those for each variable, since the choice of positions of every coordinate is done independently, i.e.

$$H^{(\varphi)} = \sum_{i=1}^{n} H^{(\varphi)}(\delta_i, m_i)$$

The magnitude $H^{(\varphi)}$ is a quantitative measure of our lack of knowledge of which of the N possible variants corresponds to the best value of criterion $F(x)$. At the beginning of the search if there is no a priori information of preference of one value of a variable to another, this entropy is equal to

$$H^{(0)} = \sum_{i=1}^{n} H_i^{(0)} = \sum_{i=1}^{n} \log l_i = \log N$$

Let us denote the relative number of experiments carried out by $\tau_\varphi = \frac{t_\varphi}{T}$,

the number of experiments remaining by $\tau_\varphi = \tau - t_\varphi$. Let us require that the entropy should decrease steadily to zero as τ_φ is increasing to one. For that, the function $H^{(\varphi)} = \log \lambda_\varphi + H^{(0)}$ where

$$\lambda_\varphi = \frac{\kappa \tau_\varphi^2}{2} + \left[a^{-H^{(0)}} - \left(\frac{\kappa}{2} + 1 \right) \right] \tau_\varphi + 1 \quad (4)$$

is introduced.

The quantity a is on a logarithmic basis. Parameter κ determines the adaptation degree. The less κ is, the less is the degree of adaptation. This parameter value must be chosen from the interval (-2, 2) in order $0 \leq \lambda_\varphi \leq 1$.

If the quantity τ_φ is changing from 0 to 1 then λ_φ is changing from 1 to $a^{-H^{(0)}}$, entropy $H^{(\varphi)}$ is decreasing steadily from $H^{(0)}$ to 0.

The condition on which the choice of concrete value of the parameter κ depends will be given below.

After carrying out τ_φ tests the entropy is decreasing in comparison with the initial one in $\frac{H^{(\varphi)}}{H^{(0)}}$ times. We consider that the entropy for each variable is also decreasing, i.e.

$$\frac{H_i^{(\varphi)}}{H_i^{(0)}} = \frac{H^{(\varphi)}}{H^{(0)}}$$

Then

$$H_i^{(\varphi)} = H^{(\varphi)} \frac{H_i^{(0)}}{H^{(0)}} = (\log \lambda_\varphi + H_0) \cdot \frac{\log \ell_i}{\log N} \quad (5)$$

Let us equate entropy $H^{(\varphi)}(b_i, m_i)$ and that $H_i^{(\varphi)}$ defining by correlations (3) and (5):

$$H^{(\varphi)}(b_i, m_i) = H_i^{(\varphi)} \quad (6)$$

Now, we determine the number of positions m_i . For that, the maximum number of the position for which one can choose even one test from the remained number of those τ_φ is determined. This number is determined thus: at first, we suppose $m_i = \ell_i$ and from equation

$$P_{\ell_i}(b_i, \ell_i) \cdot \tau_\varphi = 1$$

the coefficient value b_i is determined. Let us denote it by $b_{i\,max}$. If $H^{(\varphi)}(b_{i\,max}, m_i) \geq H_i^{(\varphi)}$, then m_i is supposed to be equal to $\ell_i - 1$. A new value m_i is tested in the same way. This procedure is repeated until the inequality is fulfilled

$$H^{(\varphi)}(b_{i\,max}, m) \leq H_i^{(\varphi)} \quad (7)$$

The number m_i is the first one m for which the inequality (7) is fulfilled.

If the number m_i is defined, then the coefficient b_i is determined from equation (6) with the method of dichotomy.

When b_i and m_i have been determined, one can compute probabilities $P_\beta(b_i, m_i)$ for $\beta = 1, 2, \ldots, m_i$ using expression (1).

Further, the number $\tau^{(\varphi+1)}$ tests included in the $(\varphi+1)$-th group is determined. It must be such a number so that one can regulate the values of every variable X_i by the quantity F when the results of experiments $F_1, F_2, \ldots, F_{t\varphi}$ are obtained. The number $\tau^{(\varphi+1)}$ must be as small as possible in order to carry search as "carefully" as possible. For that, we require that the value of variable which in a given moment of search is occupying the last position should be included in a test only one time. From that

$$\tau^{(\varphi+1)} = \max_{i=1,\ldots,n} \{ окр\ \tau_i^{(\varphi+1)} \}$$

where $окр\{\tau_i^{(\varphi+1)}\}$ is the nearest integer from $\frac{1}{P_{m_i}(b_i, m_i)}$.

The test number $z_{i\beta}^{(\varphi+1)}, \beta=1, \ldots, m_i, i=1, \ldots, n$, is determined by means of expression (2).

Then, the set of $z^{(\varphi+1)}$ points in n - dimensional space of positions is chosen in correspondence with distribution law given by function of "confidence". The choice of these points is done so that the number of points corresponding to the β -th position of variable $z_{i\beta}$ should be equal to X_i . Coordinates of these points are represented as matrix $B^{(\varphi+1)}$ of $(z^{(\varphi+1)} \times n)$ dimension.

The procedure for obtaining matrix $B^{(\varphi+1)}$ is repeated several times. The matrix to which corresponds the least number of coinciding lines (coinciding points) is chosen from those obtained. If the relative number of coinciding lines exceeds a threshold then the degree of adaptation is decreased (value of parameter κ is decreased beginning with $\kappa = 2$) and the process of obtaining matrixes $B^{(1)}, B^{(2)}, \ldots, B^{(\varphi)}$ is repeated from the very beginning.

If κ is decreasing to -2 and one can not obtain all matrixes $B^{(1)}, B^{(2)}, \ldots, B^{(\varphi)}$, then we output matrixes obtained and the number of tests not distributed.

Further experimental planning involved the conversion of matrix $B^{(\varphi)}$ to $G^{(\varphi)}$ by means of vectors $\alpha^{(\varphi)}$ set and renewal $\alpha^{(\varphi)}$ on base of results obtained $F_1, F_2, \ldots, F_{z\varphi}$. The procedure for obtaining matrix $G^{(\varphi)}$ can be done without using a computer.

REFERENCES

1. Суппес П., Зинес Дж. Основы теории измерений. Сб."Психологические измерения", Москва, изд."Мир",1967г.
2. Корбут А.А.,Финкельштейн Ю.Ю. Дискретное программирование. Москва, изд. "Наука", 1969 г.
3. Лбов Г.С. Выбор эффективной системы зависимых признаков. "Вычислительные системы", вып.19, Новосибирск, 1965.
4. Бусленко Н.П., Шрейдер Ю.А. Метод статистических испытаний. Москва, ФМ, 1961.