Phase Intervals and Genetic Algorithms Based Competition Task Solution

Serhiy Shtovba
Vinnitsa State Technical University
Khmelnitskoe Shosse, 95, 21021, Vinnitsa, Ukraine
Phone: +380-432-440430, Fax: +380-432-465772
e-mail: shtovba@svitonline.com

ABSTRACT: This paper presents the competition task solution based on the use of phase interval method and genetic algorithm of optimization. Difficulties and peculiar of the competition task are discussed also.

KEYWORDS: Prediction, Phase Interval, Genetic Algorithms.

1. FORMALIZED STATEMENT OF THE PROBLEM

The competition problem is a typical example of the pattern recognition task. From mathematical point of view the design of competition task pattern recognition model consists in finding out of mapping F of source information vector $X=\{x_1, x_2, ..., x_{85}\}$ into two-element set $Y=\{1,0\}$. For model design can be used only the set of experimental data: (X_r, y_r), $r=\overline{1,5822}$. The use of other knowledge is impossible.

2. PREDICTION MODEL

The competition problem solving is large dimension identification task. In connective with this we must choose such kind of model, which provide the reasonable time of design under good prediction condition. Design time for model based on universal approximater, such as neural network, is more than for model, structure of which is determined by the problem. Intelligent models with unique structure are fuzzy inference systems and neural-fuzzy networks. Structure of ones is determined by fuzzy rules. As far as we know the extraction rules from data automatically is possibly only for small scale problem. For the competition task it is hardly, because of number of rules will be huge. The use of a hierarchical knowledge bases allow to decrease fuzzy rules number, but, unfortunately, we haven't any information about techniques for automatic creation of hierarchical logic evidence tree.

Our prediction model was designed on base of the phase interval method [2]. Its main idea consists in selection such class for the object X, which has minimal distance between the center of class and the object. In mathematical notes it is:

$$y=\begin{cases}1, \text{if } d\_yes(X)<d\_no(X) \\ 0, \quad\quad\quad \text{otherwise,}\end{cases}$$

where y - class corresponding to object X ($y=F(X)$, $y\in Y$);

$d\_yes(X)$ ($d\_no(X)$) - minimal distance between center of class 1 (0) and object X. These distances to calculate by the way:

$$d\_yes(X)=\min_{j=\overline{1,k}} \sum_{i=\overline{1,85}} w_i \cdot \left|x_i - x_{i,j}^1\right|,$$

$$d\_no(X)=\min_{t=\overline{1,n}} \sum_{i=\overline{1,85}} w_i \cdot \left|x_i - x_{i,t}^0\right|,$$

where w_i - weigh of the i-th coordinate of space;

k (n) - number of different centers of the class 1 (0);

$x_{i,j}^1$ ($x_{i,t}^0$) - the i-th coordinate of the j-th (t-th) center of class 1 (0).

Main advantages of the model are:

- simplicity;
- possibility of draft parameters of models putting from source data directly;
- possibility of identification by genetic algorithms.

In this case the identification may be formulates as finding of such vectors $W=\{w_1,w_2,...,w_{85}\}$, $X_j^1=\{x_{1,j}^1,x_{2,j}^1,...,x_{85,j}^1\}$ ($j=\overline{1,k}$) and $X_t^0=\{x_{1,t}^0,x_{2,t}^0,...,x_{85,t}^0\}$ ($t=\overline{1,n}$), which provide the best prediction:

$$\sum_{r=1,5822}\left|y_r-F(X_r,W,X_j^1,X_t^0)\right|\to\min,\quad j=\overline{1,k},\quad t=\overline{1,n}.$$

For solving of the above pointed optimization task we used genetic algorithms [1]. In this case a decision variant represents as the next chromosome:

W	X_1^1	...	X_k^1	X_1^0	...	X_n^0

where each cell consists of 85 genes. For initial population the centers of classes 1 and 0 was selected randomly from source experimental data set. It is allows to reduce optimization time.

3. RESULTS

Results of prediction by several models are showed in table 1. All models has same number of center of classes: $k=8$ and $n=7$.

Number of model	Correctness of prediction on the training set, %	Score
1	94.06	77
2	94.04	75
3	89.59	76
4	71.95	89

Table 1: Prediction results

Optimal weighs allow to point the next most important factors:

3 Average size household;
10 Married;
15 Household with children;
22 Middle management;
31 Home owners;
32 1 car;
35 National Health Service;
42 Average income;
44 Contribution private third party insurance;
47 Contribution car policies;
59 Contribution fire policies;

and noise factors:

49 Contribution motorcycle/scooter policies;
52 Contribution tractor policies;
53 Contribution agricultural machines policies;
60 Contribution surfboard policies;
61 Contribution boat policies;
70 Number of motorcycle/scooter policies;
75 Number of moped policies;
82 Number of boat policies.

The centers of class 1 (0) may be interpreted as examples of typical customers, who have (don’t have) got the mobile home policies.

3. GENERAL DISCUSSION

The compare of the solutions of participants and prediction by at random is showed on Fig.1. The solid line is score distribution for case of random guess, the bars are distribution of participant scores. The Fig.1 shows intelligent techniques to increase efficiency of decision making more than twice.

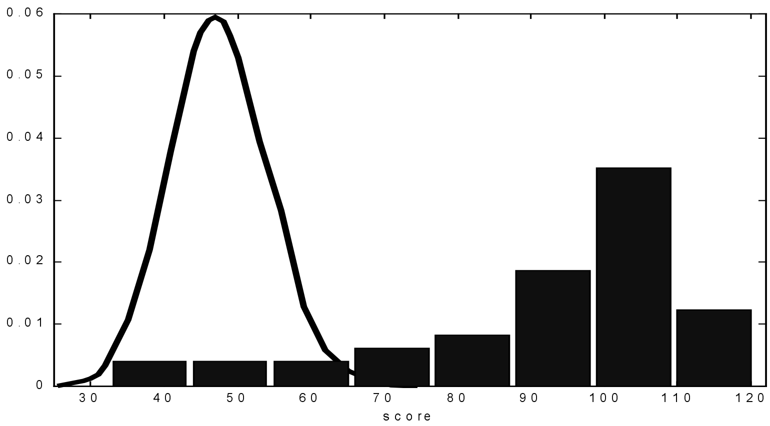

Figure 1: Score distributions

As is generally known, nobody have got the maximally possibility level of score. The best results is 121 guess from 238. Some causes of it are considered below.
The design of a mathematical model consists of the next stages:

1) problem statement;
2) experimental data collection;
3) structural identification;
4) parameter identification.

The quality of a model relates from quality execution of above pointed stages. The participants used different approachs for the third and forth stages execution, because the possibility of mistakes for this stage is low. Let's look at experimental data and statement of problem more intently.
Generally, some inconcordance between prediction levels for training and testing sets (see tabl.1, for example) may be connected with differences of the ones. In other words, there are a lot elements in testing set the distances from it to some elements of training set are long. We found the ones in the testing data. It is records number 92, 2201, 3001, 3048, 3367, 3487, 3760. The number of the ones is small, because influence of its is low.
Usually, mistakes in a problem statement are connected with errors in influence factors determination. It is including of noise factors and (or) canceling of important ones. There are a lot of noise factors (see chapter 3) in the problem statement by our opinion. But it is not seriously always, because the intelligent techniques allow to filter of some noises. Canceling of important factors is more hard error, because of any identification method don't allows to compensate it. From our point of view, some factors are not included into the problem statement. This assertion is supported up by availability of pairs of same records in testing data with different values of target parameter. They are records 181 and 3657, 279 and 1183 etceteras. There are 37 such pairs in the testing data.

ACKNOWLEDGMENTS

The author is grateful to students of Vinnitsa State Technical University Mikolu Majstrenko and Vitaly Bilousa for assistance in computational experiments.

REFERENCES

1. Gen M., Cheng R. Genetic Algorithms and Engineering Design.- John Wiley & Sons.- 1997.- 352p.
2. Rastrigin L. Computers, systems, networks. Moscow: Nauka Publisher House.- 1982.- 223p. (In Russian).