THE SOLUTION OF PROBLEM OF PARAMETERIZATION OF THE PROXIMITY FUNCTION IN ACE USING GENETIC ALGORITHM

Khamroev Alisher

Senior researcher, Centre for development software and hardware program complexes at the Tashkent University of Information Technologies, Tashkent, Uzbekistan

Abstract

In this work, a new approach for defining the value of the proximity function, which is carried out in the second step of the Algorithms for Calculating Estimates (ACE) in the area of Pattern Recognition, is presented. The value of the proximity function is defined as a part of corresponding features of two objects. The main attention is paid to essential features of the polytypic in a given training set. One of the important problems of the ACE is to compare the values of fuzzy attributes. The main idea of this approach is considering the proximity the corresponding quantitative and qualitative features together. Here a complexity of comparing the qualitative features and an approach of overcoming such complexity are considered. Such features include the features with fuzzy values. The membership function of fuzzy set theory is used for determining membership degrees of the feature values describing with linguistic values for improve the quality of ACE. The steps of the algorithm for transfer the results is obtained from the comparison of the two values of fuzzy feature by using membership function to the proximity function. The membership function with two parameters (b and c) is used. For defining optimal values of these parameters evolutionary algorithms for solving optimization problems are used, one of them is Genetic algorithm. By using genetic algorithm initial parameters' values of the membership function are generated and transmitted to the proximity function. The ACE is run and value of functional quality is defined during the training process with given training set. If the value of the functional quality is not sufficiently high than the values obtained by Genetic algorithm, these values are regenerated using special operators (selection, crossover, mutation) of the Genetic algorithm. The algorithm for selection optimal values of the parameters of the membership function using the Genetic algorithm is given.

Key Words: *ACE, proximity function, Genetic algorithm, membership function, parameters, operators.*

----------***----------

1. INTRODUCTION

Problem of recognition is connected with the study of human creative activity, the study of how the processes of making a decision on solving various problems, how targets are selected and concepts are formulated occur in the human brain. These questions lie in sphere the so-called heuristic principles and methods for solving problems. The task of recognition of objects of different physical nature is that in terms of set of features, which show an unknown object from a defined class, it is needed to determine to what specific object out of this class this particular set of features corresponds to.

For the solution of the problem of recognition based on the precedents many methods and algorithms were developed. One of such recognition algorithms are algorithms based on the principle of partial precedent (the algorithms for calculation estimates - ACE), developed and improved under the supervision of Yu.I. Zhuravlev [1, 2]. At the core of these models is the idea of finding the right partial precedents in features descriptions source data (informative fragments of feature values or representative sets).

2. STATEMENT OF THE PROBLEM

Let there be a general set of $\{S\}$ which consists of original data (objects) $S_j \in M$, $j = 1,2,...,m$. Usually, naturally the objects S_j are given with polytypic features $X_i, i = 1,2,..n$, $S_j = (x_{1j}, x_{2j}, ..., x_{nj})$, $x_{ij} \in X_i$. The elements (values of features) X_i make up an alphabetic features, for example, they can be expressed in terms of "yes/no", "yes/no/unknown", the numerical values, linguistic values, the values of a set of possible options, etc. The task of recognition – is the classification of source data to a specific class using a selection of essential features, which characterize the data, out of the total weight of the non-essential data.

2.1 The Algorithms for Calculating Estimates problem

It is known [1], that in the algorithms, the calculation of recognition is based on comparing the recognizable object with the standard objects by different sets of features using the voting procedures. In most cases, the voting procedure is conducted for quantitative features of objects. However, the features of objects can be expressed by fuzzy values (linguistic terms). In such cases, the proximity between relevant two fuzzy features of objects should be considered. For this we have to modify the second step of setting of

ACE for the purpose of correctly calculating the value of the proximity function $r(S_j, S_q)$. Here S_q - control objects.

At the first stage, ACE is formed from full set of $(X_1,...,X_n)$ under the possible collection of k featuring $\Omega_1,...,\Omega_k$ so-called system of support collection. Here k – the length of the voting under the collection of n, which represent some $\tilde{\omega}$-part of the features of vector objects.

In the second stage of ACE the proximity function $r(\tilde{\omega}S_j, \tilde{\omega}S_q)$ the value which determines measure of proximity of the parts of rows (so-called $\tilde{\omega}$-parts) is calculated. The proximity function $r(\tilde{\omega}S_j, \tilde{\omega}S_q)$ is determined using following functions $\rho(x_{i_v j}, x_{i_v q})$, which determine degree of proximity of features. These functions, as a rule, are selected in accordance with the character of used features.
The value of the proximity function can be set in the following types (comparison: a strict-b-soft):

$$1)\ r(\tilde{\omega}S, \tilde{\omega}S_q) = \begin{cases} 1; & \tilde{\omega}S = \tilde{\omega}S_q \\ 0; & \text{otherwise,} \end{cases}$$

$$2)\ r(\tilde{\omega}S, \tilde{\omega}S_q) = \begin{cases} 1; & \sum_{v=1}^{k} \tilde{\rho}(x_{i_v j}, x_{i_v q}) \le \varepsilon \\ 0; & \text{otherwise,} \end{cases}$$

here $\tilde{\rho}(x_{ij}, x_{iq})$ - comparison function between the features $x_{i_v j}$ and $x_{i_v q}$, ε - threshold for $\tilde{\omega}$-parts.

We will consider some functions, which determine the degree of proximity of the values of the corresponding feature of two objects:
a) if the features are given in binary values, the following function is entered:

$$\rho(x_{ij}, x_{iq}) = \begin{cases} 1; & x_{ij} = x_{iq}, \\ 0; & x_{ij} \ne x_{iq} \end{cases}$$

b) if the features are given in quantitative values, the following function is entered which uses the parameter ε_i - thresholds:

$$\rho(x_{ij}, x_{iq}) = \begin{cases} 1, \text{если} \mid x_{ij} - x_{iq} \mid \le \varepsilon_i, \\ 0, \text{otherwise,} \end{cases}$$

c) if the features are given in fuzzy values, fuzzy terms, linguistic values (Example: "low", "below average", "average", "slow", "fast", etc.), or so-called fuzzy sets $A_1, A_2, ..., A_t$, it is difficult to use Euclidean metric and it is needed to introduce a fuzzy metric. In this case, another function (membership function) which determines the degree of fuzzy features (term set) is selected. Each term set is described by their membership functions.

2.2 Problem Formulation Using The Membership Function

In the process of solving some problems, the standard forms of the membership function (bell-shaped, trapezoidal, triangular, Gaussian type, etc.) are used. Usually these functions are described by two or three parameters. We will consider a membership function with two parameters (e.g., b, and c) [4].

Usually a specialist of the subject area determines the appropriate types and parameters of the membership function. In addition, builds a membership function, which corresponds to the terms of the values of each feature based on the input data. However, it is not always possible to involve the experts of the subject area for solving applied problems. Therefore, there is a need to develop the algorithm of the settings (parameterization) of the proximity function based on the set provided by the specialists.

A fuzzy set is denoted by a set of pairs $\langle \mu_{A_v}(\hat{x}) / x \rangle$, where x takes some informative value, and $\mu_{A_v}(\hat{x})$ displays the x to the unit interval, taking values from 0 to 1. Here $\mu_{A_v}(\hat{x})$ represents the degree of membership of x to something.
To solve the problem of parameterization, you can use evolutionary and multi-agent methods and algorithms, such as Genetic algorithms, neural networks, the algorithm "Ant colony", Immune algorithm, etc. Each method and algorithm is characterized by its own set of properties and parameters.

We consider the use of membership functions in a fuzzy information about the features of the object, as well as consider the problem of the parameterization of the membership function using Genetic algorithm. By means of this algorithm, the optimization problem is solved with the following target (fitness) function:

$$y = F(S, B, C)$$

where S - a set of objects with fuzzy values of features; $B = (b_1, b_2, ..., b_q)$ and $C = (c_1, c_2, ..., c_q)$ - the vectors of parameters of the settings of the membership functions; q - total number of terms; F - the target function, which adopts appropriate values and determines the quality of the optimization stage after realization of step of the ACE.

Let the set under study be in the form of m pairs of experimental data:

$$(S_j, y_j),\ j = \overline{1, m}$$

where $S^j = (x_{1j}, x_{2j}, ..., x_{nj})$ – input vector and y_j appropriate values of the output variable y for jth pair <input-output>, $y_j \in [\underline{y}, \overline{y}]$.

In accordance with the method of least squares the problem of optimum settings of fuzzy model can be formulated as follows: finding the vectors (B, C), which satisfy the following restrictions

$$b_t \in [\underline{b}_t, \overline{b}_t],\ c_t \in [\underline{c}_t, \overline{c}_t],\ t = \overline{1,q},$$

and provide

$$\sum_{j=1}^{m} \left[F(S_j, B, C) - y_j \right]^2 = \min_{B,C} \qquad (1)$$

where $\underline{b}_t$ and $\underline{c}_t$ ($\overline{b}_t$ and $\overline{c}_t$) lower (upper) boundary, which gets a value of parameters of the proximity function.

3. THE SOLUTION OF THE PROBLEM

To solve the problem (1) the Genetic algorithm optimization is proposed [4].

For the realization of the Genetic algorithm the method of encoding the values of the unknown parameters B and C should be defined. We reduce the parameters B and C in a single vector:

$$\theta = (B, C) = \left(b_1^1, c_1^1, ..., b_1^{t_1}, c_1^{t_1}, ..., b_n^1, c_n^1, ..., b_n^{t_n}, c_n^{t_n}\right)$$

where t_i - the number of terms-grade of input variable x_i, $i = \overline{1,n}$, $t_1 + t_2 + ... + t_n = q$;

q – the total number of terms;

Vector θ uniquely defines a model $F(S, B, C)$.

3.1 Crossover Operator

Since the operation of crossing is a basic operation of the Genetic algorithm, its performance is primarily dependent on the performance of used crossing operations. As a result, crossing two parent chromosomes θ_1 and θ_2 provides Ch_1 and Ch_2 chromosome offspring through the exchange of genes with respect to (n)th point of the crossing.

It should be noted that because of a lot of $A_i = \{a_i^1, a_i^2, ..., a_i^{t_i}\}$ terms-estimates of the input variables are in ascending order (for example: “low”, “medium”, “high”, etc.), the introduction of cross-breeding operation may disturb this order. Therefore, after the exchange of genes control should be implanted to ensure that the set of terms were ordered. We introduce the following notation:

$b_{ip}^{\theta_1}$ – ip^{th} parameter of b in a chromosome-parent θ_1,

$b_{ip}^{\theta_2}$ – ip^{th} parameter of b in a chromosome-parent θ_2,

$b_{ip}^{Ch_2}$ – ip^{th} parameter of b in a chromosome-parent Ch_2,

$$\theta_1 = \left(b_1^{1(1)}, c_1^{1(1)}, |..., b_1^{t_1(1)}, c_1^{t_1(1)}, ..., b_n^{1(1)}, c_n^{1(1)}, |..., b_n^{t_n(1)}, c_n^{t_n(1)}\right)$$
$$\theta_2 = \left(b_1^{1(2)}, c_1^{1(2)}, |..., b_1^{t_1(2)}, c_1^{t_1(2)}, ..., b_n^{1(2)}, c_n^{1(2)}, |..., b_n^{t_n(2)}, c_n^{t_n(2)}\right)$$

After the crossover operation the following new chromosomes are yielded:

$$Ch_1 = \left(b_1^{1(1)}, c_1^{1(1)}, |..., b_1^{t_1(2)}, c_1^{t_1(2)}, ..., b_n^{1(1)}, c_n^{1(1)}, |..., b_n^{t_n(2)}, c_n^{t_n(2)}\right)$$
$$Ch_2 = \left(b_1^{1(2)}, c_1^{1(2)}, |..., b_1^{t_1(1)}, c_1^{t_1(1)}, ..., b_n^{1(2)}, c_n^{1(2)}, |..., b_n^{t_n(1)}, c_n^{t_n(1)}\right)$$

Algorithm operation crossing two parental chromosomes θ_1 and θ_2, which will result in offspring of Ch_1 and Ch_2 is as follows:

Step 1. Generate a random number z_i in an amount (n), such that $1 \le z_i < t_i$, where t_i - the number of terms-grade input variable x_i, $i = \overline{1,n}$.

Step 2. Genes are exchanged in accordance with the values found z_i exchange points on the rules:

$$b_{ip}^{Ch_1} = \begin{cases} b_{ip}^{\theta_1}, & p \le z_i, \\ b_{ip}^{\theta_2}, & p > z_i. \end{cases} \quad b_{ip}^{Ch_2} = \begin{cases} b_{ip}^{\theta_2}, & p \le z_i, \\ b_{ip}^{\theta_1}, & p > z_i, \end{cases}$$

$1 \le p < t_i, i = \overline{1,n}$.

Step 3. Control over the order of terms is set:

$$(b_{i\beta} > b_{i\eta}) \wedge (\beta < \eta) \Rightarrow b_{i\beta} \leftrightarrow b_{i\eta}, c_{i\beta} \leftrightarrow c_{i\eta},$$
$$1 \le \beta, \eta \le l_i,\ i = \overline{1,n},$$

where $\leftrightarrow$ symbol of exchange.

The mutation operator. Each element of the vector S can undergo operation of mutation probability p_m. The mutation of the element *t* is denoted by *Mu*(*t*):

$$Mu(b_{ip}) = RANDOM([\underline{x}_i, \overline{x}_i]) \qquad (2)$$
$$Mu(c_{ip}) = RANDOM([\underline{c}_i, \overline{c}_i]) \qquad (3)$$

where $[\underline{c}_i, \overline{c}_i]$ - the range of possible values of the concentration factor-stretching features proximity terms, estimates of input variable x_i, $[\underline{c}_i, \overline{c}_i] \subset (0, +\infty]$, $i = \overline{1,n}$;

3.2 Mutation Operator

The algorithm of the mutation operation will be:

Step 1. For each of the element $z \in \theta$ in the vector θ a random number $z = RANDOM([0,1])$ is generated.

If z>pm the mutation does not produce otherwise proceed to step 2.
Step 2. Carry out the operation of the mutation element $z \in \theta$ in accordance with formulas (2)-(3).
Step 3. Control over ordering is set.

3.3 The Function of Compliance

Denote the function of matching chromosomes θ in terms $FF(\theta)$ (from the English. Fitness function). As the compliance function we use the optimization criterion, taken with the minus sign.

For the developed models $F(S, B, C)$ function of ACE, matching chromosomes derived from the criterion (1), will be:

$$FF(\theta) = -\sum_{j=1}^{m}\left[F(S_j, B, C) - y_j\right]^2 .$$

The minus sign is needed so that the meaning of compliance function has not changed, that is, the lower the quality of the proposed model, the lower the value of its compliance function.

In accordance with the principles of Genetic algorithm choice of parents for the operation of crossing must be carried out not by accident. The greater the value of the function corresponds to a certain chromosome, the greater the likelihood that this will give the offspring chromosome.

A method for determining the parents based on the fact that each chromosome in the population S_i assigned a number p_i, such that:

$$p_i \geq 0, \sum_{i=1}^{K} p_i = 1,\ FF(\theta_i) > FF(\theta_j) \Rightarrow p_i > p_j,$$

K - number of chromosomes in the population. The number p_i is interpreted as the probability of which is calculated as follows:

$$p_i = \frac{FF(\theta_i)}{\sum_{j=1}^{K} FF(\theta_j)}.$$

Using a series of numbers p_i, chromosome parents for the operation of crossing we find the following algorithm:
Step 1. Postpone a number of p_i on the horizontal axis (Fig. 2).

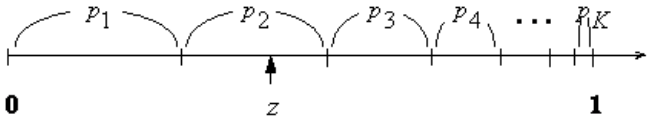

Figure 1. Selecting the the parent chromosomes

Step 2. We generate a random number z (Fig. 2) having a uniform law of distribution on the interval [0,1].
Step 3. As a parent chromosome we choose θ_i, the corresponding sub-interval p_i, which hit number z. For example, in Fig. 2 the generated number z determines as a parent chromosome θ_2.
Step 4. We repeat steps 1-3 for the determination of the second parent chromosome.

3.4 Generating A Population

Generating a population means determining the initial set of solutions (chromosomes-parents) who are at the crossing. Given initial solutions (elements of the vector S) as follows:

$$b_i^0 = RANDOM\left(\left[\underline{x}_i, \bar{x}_i\right]\right)$$
$$c_i^0 = RANDOM\left(\left[\underline{c}_i, \bar{c}_i\right]\right)$$

where $RANDOM\left(\left[\underline{\xi}, \bar{\xi}\right]\right)$ is the operation of finding a uniform distribution on the interval $\left[\underline{\xi}, \bar{\xi}\right]$ random number.

After the random baseline variants of chromosomes, they should be subject to control for the order of terms are kept.

It is assumed that the initial population has a K parent-chromosomes.

4. THE STEPS OF THE ALGORITHM

At each iteration of the Genetic algorithm population size will increase by $K \cdot p_c$ offspring chromosomes where p_c - the factor of crossing. To maintain a constant population size (K), before the next iteration the worst $K \cdot p_c$ chromosomes (in terms of matching functions) should be discarded. In view of this, the Genetic algorithm of the optimal adjustment of fuzzy model *F(S, B, C)* ACE will be as follows:

Step 1. We generate initial population.
Step 2. Find the values of the functions $FF(\theta_i)$, $i = \overline{1, K}$ compliance by implementing the steps of the ACE.
Step 3. Define $\frac{K \cdot p_c}{2}$ pairs of parent chromosomes.
Step 4. Perform the operation of crossing each pair of chromosomes of parents in accordance with the algorithm crossing.
Step 5. With probability p_m perform mutation of derived offspring chromosome according to the mutation algorithm.
Step 6. From the resulting population size of $K + K \cdot p_c$ chromosomes $K \cdot p_c$ chromosomes, which have the worst value of the function of compliance $FF(\theta_i)$, are discarded.
Step 7. If the chromosome S_i, for which $FF(\theta_i) = 0$, obtained, then it is the end of the algorithm, otherwise go to step 8.

Step 8. If the specified number of steps are not exhausted, then proceed to step 2; Otherwise, the chromosome having the greatest value of compliance $FF(\theta_i)$, received through the ACE, is a suboptimal solution. It is the end of algorithm.

Note that two fuzzy values considered close feature if and only if both of the characteristic values simultaneously belong to one term set.

$$\rho(x_{ij}, x_{iq}) = \begin{cases} 1, \mu^*_{A_v}(\hat{x}_{ij}) = \mu^*_{A_v}(\hat{x}_{iq}) \\ 0, \text{otherwise} \end{cases}$$

where

$$\mu^*_{A_v}(\hat{x}_{ij}) = \max_{v=\overline{1,t}}(\mu_{A_v}(\hat{x}_{ij}),$$

$$\mu^*_{A_v}(\hat{x}_{iq}) = \max_{v=\overline{1,t}}(\mu_{A_v}(\hat{x}_{iq}).$$

Here are all the range of x_{ij} fuzzy attribute given to one universal interval $[0, \nu]$ with the following ratios:

$$\hat{x}_{ij} = \nu \frac{x_{ij} - \underline{x_{ij}}}{\overline{x_{ij}} - \underline{x_{ij}}}$$

where $\overline{x_{ij}}, \underline{x_{ij}}$ - the interval of the variable x_{ij}, $i = \overline{1,n}$, $\hat{x}_{ij}$ - normalized sign.

After the calculation of proximity function $r(\tilde{\omega}S_j, \tilde{\omega}S_q)$, for $\tilde{\omega}$ - parts of objects S_j and S_q, are computed voice $\Gamma(S_j, K_u)$ to class K_u, $u = 1,2,..l$, by 3-5 stages ACE [1,2]:

$$\Gamma(S_j, K_u) = \frac{1}{N_u} \sum_{q=m_{u-1}+1}^{m_u} \sum_{\tilde{\omega} \in \Omega_A} r(\tilde{\omega}S_j, \tilde{\omega}S_q),$$

where $S_q \in K_u$, $q = \overline{m_{u-1}+1, m_u}$, $N_u, N_u = m_u - m_{u-1}$ – the number of objects in the class K_u.

5. CONCLUSION

In conclusion, it is worth noting that the optimization of parameters is considered as a multiextremal task. Using the traditional mathematical methods for solving these problems does not always produce the desired effect. Use of the Genetic algorithm especially using the random search is more preferred.

The considered algorithm makes it possible to build an adaptive model ACE based on the training sample, issued by subject matter experts.
It should be emphasized that the proposed GA, based on the idea of a random search, may require a lot of computation time and more resources. However, this problem can be solved using the technology of multiprocessor parallel computing.

REFERENCES

[1]. Zhuravlev YI, Kamilov MM, Tulyaganov Sh.E. Algorithms for calculating estimates and their application. -Tashkent: 1974. -124 pp. (in Russian).
[2]. Zhuravlev Y.I. Favorite scientific works. - M.: Master, 1998. - 420 pp. (in Russian).
[3]. Rothstein A.P. Intelligent identification technology, fuzzy sets, genetic algorithms, neural networks. - Vinnica: The UNIVERSUM, 1999. -320 p.
[4]. Rutkovskaya D., Pilinsky M., Rutkowski L. Neural networks, genetic algorithms and fuzzy systems. M.: Hotline - Telecom, 2006. - 452 p.

BIOGRAPHIES

He got his MSc degrees from specialty "Informatics" in Karshi State University 2003 respectively. He published several papers on the topic of Pattern Recognition, Data mining, such as fuzzy sets theory and the Algorithms for Calculating Estimates in many international conferences and journals.