Nonlinear Classification by Genetic Algorithm with Signed Fuzzy Measure

Honggang Wang, Hua Fang, Hamid Sharif, Zhenyuan Wang

Abstract—**In this paper, we propose a new nonlinear classifier based on a generalized Choquet integral with signed fuzzy measures to enhance the classification power by capturing all possible interactions among two or more attributes. A special genetic algorithm is designed to implement this classification optimization with fast convergence. Instead of using a discrete misclassification rate, the objective function to be optimized in this research is a continuous Choquet distance with a penalty coefficient for misclassified points. The numerical experiment shows that the special genetic algorithm effectively solves the nonlinear classification problem and this nonlinear classifier accurately identifies classes.**

I. Introduction

Supervised classification is a procedure of constructing a mathematical model based on a training data set and using the model to assign a categorical class label to any new sample element. Essentially, this type of classification procedure is an optimization problem and has been widely applied in the pattern recognition and decision making literature. Recently, Wang, sharif and Wang [1] proposed a new classifier based on genetic algorithms (GA) where the model is linear. In [1], the classifier estimated a hyperplane to separate the given data in the feature space for a linear model. However, in the real world, the data is most likely to be linearly inseparable. In this situation, nonlinear models are needed to enhance the classification power.

A naive assumption is that the contribution from all the attributes is the sum of the contribution from each individual attribute. This consideration usually results in a power loss in classification models. If the interaction among attributes towards the classification is non-ignorable, fuzzy measures (non-additive measures) should be considered. When the non-additive fuzzy measures are identified through the Choquet integral, the classifier becomes nonlinear[2-4,6,8].

Literature indicates that the genetic algorithm is an effective approach for finding the optimal solution of a non-linear classification problem[5-7]. The genetic algorithm is a parallel random search technique widely applied in parameterized optimization problems [9], although it has been shown that its search speed is sometimes slow. In this work, a specially designed genetic algorithm is adopted for solving the nonlinear classification problem and determining all unknown parameters in the model from the learning data.

In the following sections, we first introduce the fuzzy measure and generalized Choquet integral used in our classification, then present our new nonlinear classification model and special genetic algorithm. In the last section, we exhibit a numerical example.

II. Fuzzy Measure and Choquet Integrals

The use of the Choquet integrals with respect to a signed fuzzy measure has been shown as an efficient approach to aggregate information from attributes via a nonadditive set function. Let $X=\{x_1,...,x_n\}$ represent the attributes of the sample space *X* and **P**(*X*) denote the power set of *X*. The signed fuzzy measure μ is defined as:

$$\mu : \mathbf{P}(X) \rightarrow (-\infty,+\infty) \text{, where } \mu(\phi)=0$$

Let μ_i, *i*=1,..., 2^n-1, denote the values of set function μ

Let $f(x_1),...,f(x_n)$ denote the values of each attribute in an observation. The procedure of calculating the generalized Choquet integral is given in [5]:

$$(c)\int f\,d\,\mu=\sum_{j=1}^{2^n-1}[f(x'_j)-f(x'_{j-1})]\cdot\mu\,(\{x'_j,x'_{j+1},...,x'_n\})$$

Where $\{x'_j,x'_{j+1},...,x'_n\}$ is a permutation of $\{x_1,x_2,...,x_n\}$ such that: $f(x'_1),f(x'_2)...,f(x'_n)$ is non-decreasing ordered , i.e.,

$$f(x'_1)\le f(x'_2)\le ... \le f(x_n)$$

In [5, 6], the weighted Choquet integral with respect to a nonadditive measure is defined by

$$Y=(c)\int w\cdot f\,d\mu$$

where f is a nonnegative set function and restricted to be regular $(\mu(X)=1)$, and $w: X \rightarrow [0,1]$ with $\sum_{i=1}^{n} w(x_i)=1$.

On the basis of the weighted Choquet integral with respect to a nonadditive measure, we generalize it to a more comprehensive one, which is with respect to a nonadditive signed measure, that is, allowing the set function to take negative values and then, to be nonmonotone. Thus, a generalized weighted Choquet integral may be expressed by

$$Y=(c)\int(a+b\,f)d\mu$$

Honggang Wang, Hua Fang, Hamid Sharif are with the University of Nebraska Lincoln, NE, USA (email: hwang, hsharif, jfang2@unl.edu).

Zhenyuan Wang is with the Department of Mathematics, University of Nebraska at Omaha, USA; email: zhenyuanwang@unomaha.edu).

1-4244-1210-2/07/$25.00 ©2007 IEEE.

where singed measure μ is restricted to be regular $(\max_{A\subset x}|\mu(A)|=1)$, and $a=(a_1,a_2,...a_n)$, $b=(b_1,b_2,...b_n)$, are n-dimensional vectors satisfying $a_i\in[0,\infty), \min_i a_i=0$, $|b_i|\in[0,1], \max_i|b_i|=1$, and we can use it as a projection tools to reduce the complexity of the classification problem in n-dimensional space. We call a and b the matching vectors. They indicate the scaling and phase matching requirements of feature attributes. i.e., they are used to scale the diverse units and ranges of the feature attribute with respective dimensions such that the sign measure μ can reflect the interaction appropriately. Generally, Y depends on f nonlinearly due to the nonadditivity of μ. To be convenient $\mu(\{x_1\}),\mu(\{x_2\}),...\mu(\{x_n\}),\mu(\{x_1,x_2\}),\mu(\{x_1,x_3\}),...$ are sometimes abbreviated by $\mu_1,\mu_2,...\mu_n,\mu_{12},\mu_{13},...$ respectively.

III. A NEW NONLINEAR CLASSIFICATION MODEL

As is well known, any multi-class classification can be decomposed to be several consecutive 2-class classification problems. So, to simplify the discussion, only 2-class classification is considered in this work. We consider a 2-class Nonlinear classification problem with classes A and A'. Suppose that the learning data consistent of l sample points belonging to class A and l' sample points belonging to class A'. Also, suppose that all of these sample points have the same feature attributes $x_1,x_2,...,x_n$. Thus, the feature space is the n-dimensional Euclidian space R^n. The j-th sample point in A, denoted by s_j, is expressed as $s_j=(f_j(x_1),\ f_j(x_2),\ ...,\ f_j(x_n))$, $j=1,2,...,l$, while the j'-th sample point in A', denoted by $s_{j'}'$, is expressed as $s_{j'}'=(f_j(x_1'),\ f_j(x_2'),\ ...,\ f_j(x_n'))$, $j'=1,2,...,l'$.

Now we want to find a Choquet hyperplane, denoted by H, having equation

$$(c)\int(a+bf)d\mu-B=0 \qquad (1)$$

that can separate classes A and A' optimally under a certain criterion, signed fuzzy measure μ and real number B are unknown actually. Without any loss of generality, we can assume that all of these parameters, $\mu_1,\mu_2,...\mu_n,\mu_{12},\mu_{13},...$ and B are in $[-1,1)$. A natural idea for the criterion to determine these unknown parameters optimally is to minimize the total sum of signed distances of these learning sample points in two classes from respective side to Choquet hyperplane H (see Figure 1). From one side, the signed distance of a sample point s_j in A to H is just the signed distance from the projection of s_j, along with the direction parallel to H onto its normal line L, to the intersection of H and L, that is, equals to

$$d_j=\frac{(c)\int(a+bf)d\mu-B}{\sqrt{\mu_1^2+\mu_2^2+...+\mu_{2^n-1}^2}},$$

$j=1,2,...,l$. From the other side, the signed distance of a sample point s_j' in A' to H is just the signed distance from the projection of $s_{j'}'$, along with the direction parallel to H onto its normal line L, to the intersection of H and L, that is, equals to

$$d_{j'}'=\frac{B-(c)\int(a+bf)d\mu}{\sqrt{\mu_1^2+\mu_2^2+...+\mu_{2^n-1}^2}},$$

$j'=1,2,...,l'$. The projection parallel to H onto L is just a transformation identified by function

$$F(s)=(c)\int(a+bf)d\mu \text{ or } F(s)=(c)\int(a+bf')d\mu$$

from the feature space R^n to one-dimensional line L, that is, under this projection, any point $s_j=(f_j(x_1),\ f_j(x_2),\ ...,\ f_j(x_n))$ in the feature space has an image $(c)\int(a+bf)d\mu$, and the Choquet hyperplane H itself has an image B. Thus, the above-mentioned total signed choquet distance is

$$D=\sum_{j=1}^{l}d_j+\sum_{j'=1}^{l'}d_{j'}'$$
$$=\frac{\sum_{j=1}^{l}((c)\int(a+bf)d\mu-B)-\sum_{j'=1}^{l'}((c)\int(a+bf')d\mu-B)}{\sqrt{\sum_{i=1}^{2^n-1}\mu_i^2}}. \qquad (2)$$

In this formula, the Choquet distance for those misclassified points will have a negative value. As for the optimality of Choquet hyperplane H, we can imagine that H should locate with a suitable direction and be pushed by the sample points in class A forwards one side as far as possible and by the sample points in class A' forwards another side as far as possible. In case there is a gap between classes A and A', the choquet hyperplane H as the classifying boundary should pass the feature space along with the gap. This means that the total signed choquet distance D in (2) should be maximized. Such a criterion for determining the optimal hyperplane looks good. Unfortunately, it does not work well actually. In fact, if in the learning data set one class is larger than another, say $l>l'$, then class A have more power than class A' to push hyperplane H to its opposite side infinitely such that the optimization problem has no solution. Thus, we must revise above optimization model.

The revision can be realized by using a large penalty coefficient to each misclassified sample point. Let

$$c_j = \begin{cases} c & \text{if } (c)\int (a+bf)d\mu < B \\ 1 & \text{otherwise} \end{cases},$$

for $j = 1, 2, ..., l$, and

$$c_{j'}' = \begin{cases} c & \text{if } (c)\int (a+bf')d\mu > B \\ 1 & \text{otherwise} \end{cases},$$

for $j' = 1, 2, ..., l'$, where $c > |l - l'|$ is a penalty coefficient and is taken as $|l - l'| + 1$ usually. Then, a penalized total signed distance is defined as

$$D = c_j \sum_{j=1}^{l} d_j + c_{j'}' \sum_{j'=1}^{l'} d_{j'}'$$
$$= \frac{\sum_{j=1}^{l} c_j ((c)\int (a+bf)d\mu - B) - \sum_{j'=1}^{l'} c_{j'}'((c)\int (a+bf')\, d\mu - B)}{\sqrt{\sum_{i=1}^{2^n-1} \mu_i^2}}. \tag{3}$$

Thus, for given learning sample data set with two classes, the unknown parameters a, b, u and B of hyperplane H as the classifying boundary can be determined by maximizing the penalized total distance D_c in expression (3). After determining the classifying boundary H that have equation (1), for any new sample $s_l = (f_l(x_1), f_l(x_2), ..., f_l(x_n))$, if

$$(c)\int (a+bf)d\mu \geq B,$$

then we classify s into class A; otherwise, classify s into class A'

IV. A GENETIC ALGORITHM

A specially designed genetic algorithm is used to solve the optimization problem shown in section III. The components of the algorithm are listed and explained as follows.

(a) Coding and decoding.
Unknown parameters $\mu_1, \mu_2, ...\mu_n, \mu_{12}, \mu_{13}, ...$ and B a, b are coded as binary genes $g_1, g_2, ..., g_{2^n+2n}$ respectively. Thus, each gene is a bit string. The length of the bit string depends on the required precision for the solution. For example, if the required precision is 10^{-3}, then each gene consists of ten bits. Once the genes are generated in some step in the genetic algorithm, they are decoded by formulas $\mu_i = 2(g_i - 0.5)$ for $i = 1, 2, ..., 2^n - 1$, $B = 2(g_{2^n} - 0.5)$, $a = 2(g_i - 0.5)$ for $i = 2^n + 1, ..., 2^n + n$, $b = 2(g_i - 0.5)$ for $i = 2^n + n + 1, ..., 2^n + 2n$.

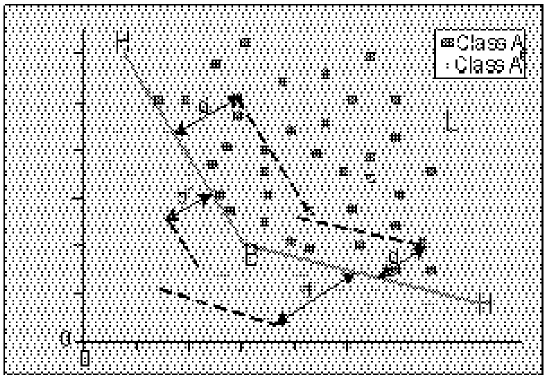


Fig. 1. 2-Dimensional data set projection based on choquet integrals

(b) Population and chromosomes.
Each chromosome is a gene string, $(g_1, g_2, ..., g_{2^n+2n})$. The population, P, consists of a large number of chromosomes. This number is called the size of the population and denoted by p. The default value of the p is 100.

(c) Chromosomes' fitness.
For each chromosome $(g_1, g_2, ..., g_{2^n+2n})$, after decoding the genes, we may obtain the values of $\mu_1, \mu_2, ...$ and a, b, B. They represent a hyperplane H via equation (1). Then, based on the given learning data, the corresponding penalized total signed Choquet distance D_c from the sample points in the data set to the hyperplane H can be calculated by (3). The relative fitness of this chromosome in the current population is defined by

$$F = \frac{D_c - D_{\min}}{D_{\max} - D_{\min}} \tag{4}$$

where

$$D_{\min} = \min\{D_c(k) \mid k = 1, 2, ..., p\}$$

and

$$D_{\max} = \max\{D_c(k) \mid k = 1, 2, ..., p\},$$

in which $D_c(k)$ is the penalized total signed distances from the sample points in the data set to the choquet hyperplane $H(k)$ corresponding to the k-th chromosome in the current population.

(d) Parents selection.
Denoting the fitness of the k-th chromosome in the current population by $F(k)$, we assign probability

$$p_k = \frac{F(k)}{\sum_{k=1}^{p} F(k)}$$

to the k-th chromosome, $k = 1, 2, ..., p$. According to the probability distribution $\{p_k \mid k = 1, 2, ..., p\}$, select two chromosomes from the population as the parents via a random switch.

(e) Producing new chromosomes.
According to the preset two-point probability distribution $(\alpha, 1-\alpha)$, via a random switch choose an genetic operation from mutation and crossover and, then, produce two new chromosomes. Repeat this procedure for $\dfrac{p}{2}$ times to get p new chromosomes.

(f) Renew population.
Calculate the total signed distance of each new chromosome and add these p chromosomes in the current population. According to the total signed distance of these $2p$ chromosomes, delete p worst from them and, then, form a new generation of the population.

(g) Stopping controller.
Repeat above procedure to get the population generation by generation until the largest penalized total signed distance, which is associated with the best chromosome in the population, has not been significantly improved for m (with default value 10) consecutive generations. Here, "has not been significantly improved" means that the improvement Δ is less than $10^{-4} \cdot d(A, A')$, where $d(A, A')$ is the distance between the centers of class A and class A' in the learning data set.

(h) After stopping, find the best chromosome in the last generation of population. Then output the corresponding values of parameters $\mu_1, \mu_2, ...\mu_n, \mu_{12}, \mu_{13}, ...$ and a, b, B.

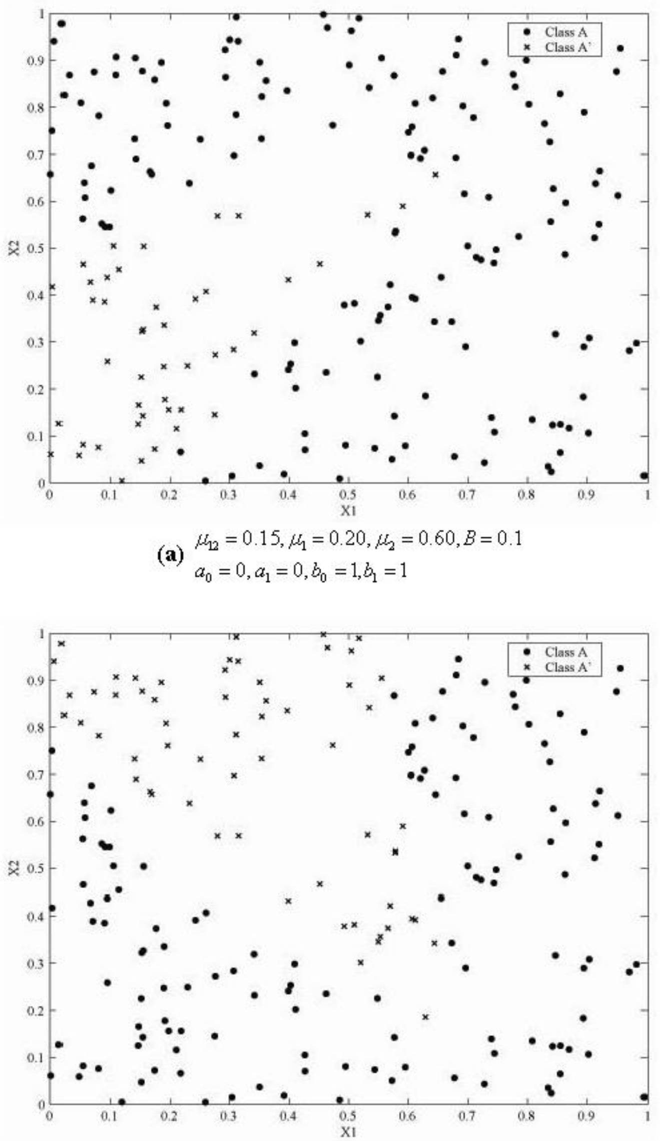
V. SIMULATIONS

We have implemented the algorithm shown in section 3 by using Microsoft Visual C++. All the functions are encapsulated into the CGenetic and CChoquet Class. Based on a training data set, the simulation runs on the WindowXP platform and regular PC desktop with AMD 1.6GHZ CPU and 512 M memory. It takes just a few minutes to stop and get an expected result.

The 2-dimensional training data set are generated by a random number generator and are separated into two classes by two straight line

$$(c)\int (a+bf)\,d\mu - B = 0$$

where μ_{12}, μ_1, μ_2 and B are pre-assigned separately, Each sample point (x_1, x_2) is labeled with class A if $(c)\int (a+bf)\,d\mu \ge B$; otherwise, (x_1, x_2) is labeled with class A'. In this way, 200 sample points are generated and labeled. Among them, in Fig 2(a) class A has 155 points, while class A' has 45 points. In Fig 2(b) class A has 140 points, while class A' has 60 points. The distribution of these sample points are illustrated in Fig 2.



(a) $\mu_{12} = 0.15, \mu_1 = 0.20, \mu_2 = 0.60, B = 0.1$
$a_0 = 0, a_1 = 0, b_0 = 1, b_1 = 1$

(b) $\mu_{12} = 0.15, \mu_1 = 0.60, \mu_2 = 0.20, B = 0.12$
$a_0 = 0.2, a_1 = 0.85, b_0 = 0.85, b_1 = -0.60$

Fig. 2. classified training data set

Running the program of our classifier based on above data set, we get the consecutive simulation results shown in Table 1 and Table 2 where G is the number of generations that have been created in the training procedure. The program for scenario in Fig 2(a) stops at the 50th generation , the scenario in Fig 2(b) stops at the 30th generation

In Table III, the second column is the number of sample points that have been correctly classified in class A by the temporary best boundary obtained in that generation, while the third column is the number of sample points that have been correctly classified in class A'. The 4-7th columns are the values of parameters μ_{12} ,μ_1 , μ_2 ,B corresponding to one of the best chromosomes in each generation. The 8th Column are penalized total choquet signed distances from the sample points in the data set to the hyperplane corresponding to one of the best chromosomes in each generation as mentioned in section III.

In Table IV, The 4-11th columns are the values of parameters μ_{12} ,μ_1 , $\mu_2, B, a_0, a_1, b_0, b_1$, The 12th Column are penalized total choquet signed distances from the sample points in the data set to the hyperplane corresponding to one of the best chromosomes in each generation as mentioned in section III.

TABLE I

Numbers of classified sample data for scenario in Fig 3(a)

Class	A	A'
Classified in A	155	0
Classified in A'	0	45

TABLE II

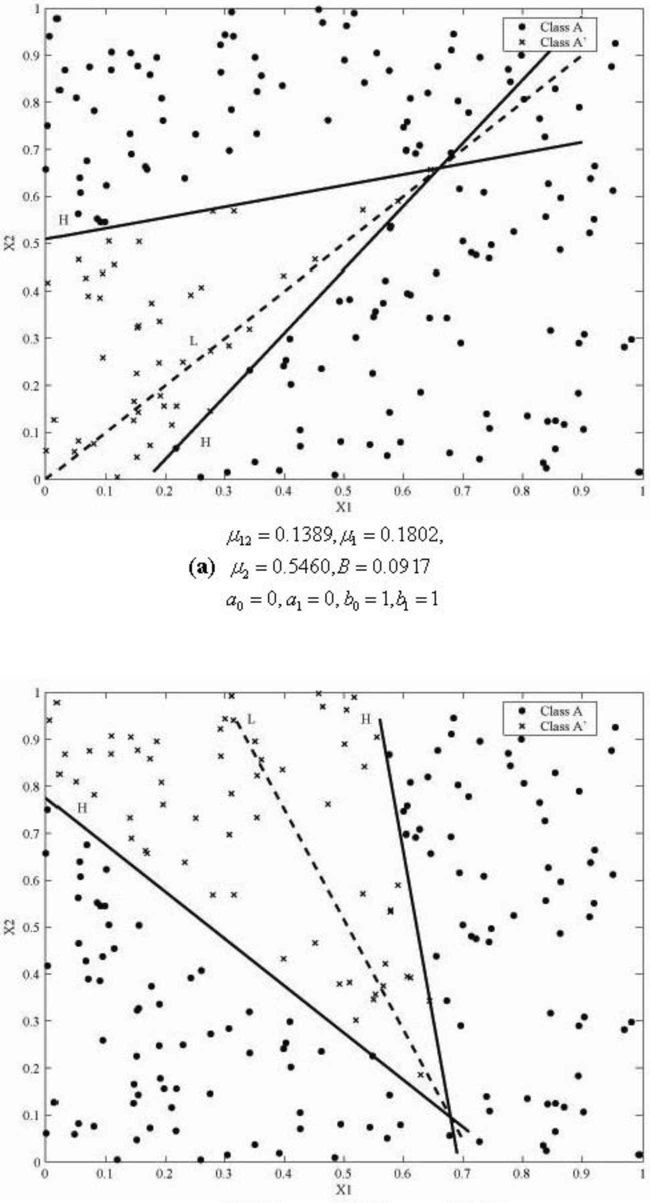
Numbers of classified sample data for scenario in Fig 3(b)

Class	A	A'
Classified in A	140	0
Classified in A'	0	60

Form Table III, in the 33th generation, the classifier has found a good chromosome whose corresponding classifying boundary can classify the training data without any misclassification. However, according the stopping condition, the program does not stop until the counter w of the stopping controller reaches 10. In Table IV, the program stops at 30th for Fig. 3(b) scenario. We can see that the values of the parameters in both Table III and Table IV almost do present the choquet hyperline as same as the pre-assigned values presentations. Table I and II summarizes the final result that shows no misclassified sample point. The classifying boundary found in the last generation is also shown in Figure 4.

VI. Conclusion

The paper addresses the classification as an optimization problem that is formed by continuous Choquet distance with a penalty coefficient for misclassified sample points instead of the misclassification rate, which is discrete [1, 8]. The experiment shows that the specially designed genetic algorithm can effectively solve the Nonlinear classification problem. By using the Choquet integral with respect to a signed fuzzy measure that describes the interaction among feature attributes, we accurately identified classes. In future, more techniques are needed to address multi-class nonlinear classification problem.



$\mu_{12} = 0.1389, \mu_1 = 0.1802,$
(a) $\mu_2 = 0.5460, B = 0.0917$
$a_0 = 0, a_1 = 0, b_0 = 1, b_1 = 1$

$\mu_{12} = 0.3830, \mu_1 = 0.6683, \mu_2 = 0.5713,$
(b) $B = 0.2633, a_0 = 0.4420, a_1 = 0.7021,$
$b_0 = 0.3614, b_1 = -0.154$

Fig. 3: Classified training data set

Appendix

Table III and Table IV.

References

[1] H. Wang, H. Sharif, and Z. Wang, A new classifier based on genetic algorithm, Proc. IPMU 2006, 2479-2484.

[2] G. Choquet. Theory of capacities. Annales de l'Institut Fourier, 5, 1953

[3] D. Denneberg. Non-Additive Measure and Integral. Kluwer Academic, 1994

[4] Z. Wang and G. J. Klir, Fuzzy Measure Theory, Plenum Press, New York, 1992.

[5] Z. Wang, A new model of nonlinear multiregression by projection pursuit based on generalized Choquet integrals, Proc. FUZZ-IEEE2002, 1240-1244, Hawaii, 2002.12

[6] K. Xu, Z. Wang, P.-A. Heng, and K.-S. Leung, Classification by nonlinear integral projections, IEEE T. Fuzzy Systems 11, No. 2 (2003), 187-201.

[7] E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, 1989.

[8] M. Liu and Z. Wang, Classification using generalized Choquet integral projections, Proc. IFSA 2005, 421-426.

[9] M. Mitchell, X. Melanie, An Introduction to Genetic Algorithms, Cambridge, Mass., MIT Press, 1996.

TABLE III
The position of generations in training process for the scenario in Fig. 3(a)

G	A	A'	μ_{12}	μ_1	μ_2	B	D
1	150	40	0.1451	0.2369	0.5423	0.1021	-2.7990
2	150	40	0.1451	0.2369	0.5423	0.1021	-2.7990
3	153	42	0.1981	0.2244	0.5687	0.1189	0.6311
				……			
32	154	44	0.1410	0.1825	0.5501	0.0927	27.2005
33	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
34	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
35	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
36	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
37	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
38	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
39	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
40	155	45	0.1387	0.1800	0.5453	0.0916	27.4156
41	155	45	0.1388	0.1802	0.5453	0.0917	27.4177
42	155	45	0.1388	0.1802	0.5453	0.0917	27.4177
43	155	45	0.1388	0.1802	0.5453	0.0917	27.4177
44	155	45	0.1388	0.1802	0.5453	0.0917	27.4177
45	155	45	0.1388	0.1802	0.5453	0.0917	27.4177
46	155	45	0.1388	0.1802	0.5456	0.0917	27.4187
47	155	45	0.1389	0.1802	0.5456	0.0917	27.4189
48	155	45	0.1389	0.1802	0.5456	0.0917	27.4189
49	155	45	0.1389	0.1802	0.5460	0.0917	27.4211
50	155	45	0.1389	0.1802	0.5460	0.0917	27.4211

TABLE IV
The position of generations in training process for the scenario in Fig. 3(b)

G	A	A'	μ_{12}	μ_1	μ_2	B	a_1	a_2	b_1	b_2	D
1	133	11	0.5685	0.5825	0.5868	0.4409	0.7617	0.1982	0.0471	0.4197	-751.6648
2	140	31	0.4899	0.5999	0.6256	0.2563	0.3893	0.5641	0.2501	-0.0771	-318.9425
3	140	31	0.4899	0.5999	0.6256	0.2563	0.3893	0.5641	0.2501	-0.0771	-318.9425
4	123	58	0.3846	0.6051	0.4342	0.2388	0.3228	0.6245	0.4720	-0.1708	-154.2835
5	135	51	0.3740	0.8851	0.7594	0.2460	0.5222	0.6800	0.1758	-0.1429	-112.1395
						……					
22	140	60	0.3845	0.6764	0.5648	0.2676	0.4536	0.7120	0.3569	-0.1580	5.9200
23	140	60	0.3845	0.6764	0.5648	0.2676	0.4536	0.7120	0.3569	-0.1580	5.9200
24	140	60	0.3845	0.6764	0.5648	0.2676	0.4536	0.7120	0.3569	-0.1580	5.9200
25	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296
26	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296
27	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296
28	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296
29	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296
30	140	60	0.3830	0.6683	0.5713	0.2633	0.4420	0.7021	0.3614	-0.1544	5.9296