International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:4, No:2, 2010

GA Based Optimal Feature Extraction Method for Functional Data Classification

Jun Wan, Zehua Chen, Yingwu Chen, and Zhidong Bai

***Abstract*—**Classification is an interesting problem in functional data analysis (FDA), because many science and application problems end up with classification problems, such as recognition, prediction, control, decision making, management, etc. As the high dimension and high correlation in functional data (FD), it is a key problem to extract features from FD whereas keeping its global characters, which relates to the classification efficiency and precision to heavens. In this paper, a novel automatic method which combined Genetic Algorithm (GA) and classification algorithm to extract classification features is proposed. In this method, the optimal features and classification model are approached via evolutional study step by step. It is proved by theory analysis and experiment test that this method has advantages in improving classification efficiency, precision and robustness whereas using less features and the dimension of extracted classification features can be controlled.

***Keywords*—**Classification, functional data, feature extraction, genetic algorithm, wavelet.

I. Introduction

WITH the development of data collection and storage technology, more and more functional data (FD) are generated in the fields of industry control, information management, Internet and simulation experiment. These kinds of information are often in the form of long time series, continuous factors depending, etc. Much potential information contained in the FD and the successful cases of the Functional Data Analysis (FDA) have suggested its advantage.

Moreover, classification is an important branch of statistics application. Many scientific and real questions end up with a classification problem, such as recognition, prediction, control, decision making and management. An observation is usually a collection of numerical measurements represented by a d -dimensional vector. However, in many real-life problems, input data are in fact (sampled) functions rather than standard high dimensional vectors, and this casts the classification problem into the class of FDA [1]. Functional data classification (FDC) is one of two common goals in application of FDA [2].

There are two barriers to handle functional situation using classical methods: the high dimension and the correlation. It is a key problem of FDA, including FDC, to reduce dimensions and correlation of FD simultaneously whereas keeping its functional features, such as integrality and smoothness. More and more studies show that wavelet-based methods are suitable to solve the problem above as the nice properties of wavelet: multi-scale time-frequency decomposition, smoothness, orthogonality, vanishing moments [1], [3]-[5], etc.

Shrinkage method [1] is popular for dimension reduction in wavelet based FDC. The shrinkage method presents good performance to keep global characters and denoise in low-dimension FD representation. It reduces the infinite dimension of the observations by considering only the first coefficients, with large power, of the data expanded on an appropriate wavelet basis. However, the aim of feature extraction for discriminant is to minimize the misdiscriminant ratio via supervised learning, which is not concerned in the shrinkage method. Some of the discarded features with small power may be non-trivial discriminatory and some of the reserved features are useless for classification. To extract optimal features according to specific problem (e.g., classification or decision based on low dimension representation) will benefit on the effect and precision of solving these problems.

In this paper, a novel automatic method using Genetic Algorithm (GA) to extract classification features from wavelet coefficients of FD is proposed, which combined GA and classification algorithm together. The optimal features and classification model are approached via evolutional study step by step in this method. It is proved by theory analysis and experiment test that this method has advantages in improving classification efficiency, precision and robustness whereas using less features and the dimension of extracted classification features can be controlled.

II. Problem Definition and Background

A. Basic Definition and Hypothesis

The problem of classification (pattern recognition or discrimination) is about guessing or predicting the unknown

This research is supported by China Scholarship Council.

Jun Wan was a visiting student of Department of Statistics and Applied Probability, National University of Singapore (NUS), 117543, Republic of Singapore. He is now pursuing the PhD degree in Department of Management Science and Engineering, National University of Defense Technology (NUDT), Changsha,410073 China (phone: +8615874837622; e-mail: wanjun_1210@hotmail.com).

Zehua Chen is with the Department of Statistics and Applied Probability, NUS, 117543, Republic of Singapore (e-mail: stachenz@nus.edu.sg).

Yingwu Chen is with the Department of Management Science and Engineering, NUDT, Changsha,410073 China (e-mail: ywchen@nudt.edu.cn).

Zhidong Bai is with the Department of Statistics and Applied Probability, NUS, 117543, Republic of Singapore (e-mail: stabaizd@leonis.nus.edu.sg).

class of an observation. An observation is a collection of measurements represented by functional data in the field of FDA.

Data are named to be functional means there is a potential function x giving rise to the observed data.

Def1 : Functional Data (FD)

A functional variable χ takes values in an infinite dimensional space. An observation x of χ is called a FD [6]. In practice, FD are usually observed and recorded discretely as *n* pairs (t_j, y_j), denoted by X, and y_j is a snapshot of the function at time t_j, possibly blurred by observational error or noise described as follows:

$$y_j = x(t_j) + \varepsilon_j$$

where the term ε_j denotes noise, disturbance, error, perturbation or otherwise exogenous which contributes a roughness to the raw data.

In general, a collection or sample of FD is concerned in practice, rather than just a single function *x*. Specifically, the record or observation X_i of the function x_i might consist of $(t_{ij}, y_{ij}), j = 1, 2, \cdots, n_i$. The argument values t_{ij} may take the same values or vary from record to record. Similarly, the interval T over which data are collected may also varies from record to record. However, these inconsistent problems can be handled using corresponding method in FDA. It is thereby assumed that t_{ij} do not vary from different records in this paper. Normally, the construction of the functional observations x_i using the discrete data y_{ij} observed separately or independently for every record i.

There are two categories in classification problem: the dual-class problem and multi-class problem. As the multi-class one can be translated into dual-class problem, only dual-class problem is discussed in this paper.

Def2 : Dual-Value Functional Data Classification

Given F is some abstract Hilbert space, and keep in mind $F = L_2([0,1])$ (i.e., the space of all square integrable functions on $[0,1]$) will be a leading example throughout the paper. The data consist of a sequence of $n+m$ i.i.d. random variables on $F \times \{0,1\}$, denoted by $\{(X_i, Y_i)\}_{i=1}^{n+m}$, where X_i's are the observations and Y_i's are the labels. Note that the data are usually artificially grouped into two independent sequences, the training sequence of length n, and the testing sequence of length m.

Def3 : Classification Rule (CR)

A Classification rule is a (measurable) function $g: F \times (F \times \{0,1\})^{n+m} \to \{0,1\}$. It classifies a new observation $x \in F$ as coming from class $g(x, (X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m}))$, denoted by $g(x)$ for the sake of convenience.

Def4 : Bayes Probability of Error (BPE)

The probability of error of a given rule g is $L_{n+m}(g) = P\{g(X) \neq Y \mid (X_1, Y_1), \cdots, (X_{n+m}, Y_{n+m})\}$, where (X, Y) is independent of the data sequence and is distributed as (X_i, Y_i) [1].

B. Wavelet-Based Functional Representation via Features

Functional representation is the process to represent the observations $\{(t_{ij}, y_{ij})\}_{j=1}^{n_i}$ of x_i in the form $y = f(t)$ in FDA. Basis function procedures usually represent a function $f(t)$ by a linear expansion in terms a series of known basis functions $\phi_v(t)$, i.e.,

$$f(t) = \sum_v a_v \phi_v(t). \quad (1)$$

Functional representation is actually a process of smooth fitting, which is convenient for FD reduction whereas keeping functional characters such as continuity. The coefficients $\{a_v\}$ character the information of functional data corresponding to different basis functions $\{\phi_v\}$. It is important to extract features effectively for classification problem, because it will impact on the FD reduction and classification.

The most popular basis systems are spline basis, Fourier basis and wavelet basis. High dimension and high correlation are correlative characters of FD which are also the difficult problems that should be resolved in FDA. A standard answer to both problems of FD is to extend PCA [7] or ICA [8] method as well as to extend wavelet method [1], [9]. Wavelet-based methods solve both of the problems simultaneously and automatically. Additionally, they are computationally faster and automatically adapt to spatial and frequency inhomogeneities of the FD. Therefore, wavelet basis is used for representation and feature extraction in this paper.

Wavelet based function fitting is also named wavelet transform or decomposition. Wavelet basis can be constructed by dilate and translate the scaling function and mother wavelet function [10]. Given wavelet function $\varphi(t)$, a series of orthonormal basis can be formed to represent a signal function $f(t) \in L^2(\mathbb{R})$ as follow:

$$f(t) = \sum_{k \in \mathbb{Z}} c_{L,k} \phi_{L,k}(t) + \sum_{j \geq L} \sum_{k \in \mathbb{Z}} d_{j,k} \varphi_{j,k}(t), \quad (2)$$

where $\mathbb{Z}$ is the set of all integers $\{0, \pm 1, \pm 2, \cdots\}$, the coefficients $c_{L,k} = \int f(t) \phi_{L,k}(t) dt$ are considered as the coarser-level coefficients characterizing smoother data patterns, and $d_{j,k} = \int f(t) \varphi_{L,k}(t) dt$ are viewed as the

finer-level coefficients describing (local) details of data patterns. In practice, the following finite version of the wavelet series approximation is used:

$$\tilde{f}(t)=\sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t)+\sum_{L\le j<J}\sum_{k\in\mathbb{Z}} d_{j,k}\varphi_{j,k}(t), \quad (3)$$

where $J>L$ and L correspond to the coarsest resolution level.

Consider a sequence of data $\mathbf{y}=(y(t_1),\cdots,y(t_N))'$ taken from $f(t)$ or obtained as a realization of $y(t)=f(t)+\varepsilon_t$ equally spaced discrete time points $t=t_i's$, where $\varepsilon_t's$ are independent and identically distributed (i.i.d.) noises. The discrete wavelet transform (DWT) of $\mathbf{y}$ is defined as $\mathbf{d}=\mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is the orthonormal $N\times N$ DWT-matrix. According to (3), the coefficients are denoted by $\mathbf{d}=(\mathbf{c}_L,\mathbf{d}_L,\mathbf{d}_{L+1},\cdots,\mathbf{d}_J)$, where $\mathbf{c}_L=(c_{L,0},\cdots,c_{L,2^L-1})$, $\mathbf{d}_L=(d_{L,0},\cdots,d_{L,2^L-1}),\quad \mathbf{d}_J=(d_{J,0},\cdots,d_{J,2^J-1})$ are called scales or subbands. Using the inverse DWT, the $N\times 1$ vector $\mathbf{y}$ of the original signal curve can be reconstructed as $\mathbf{y}=\mathbf{W}'\mathbf{d}$. The process of transforming a data set via the DWT closely resembles the process of computing the Fast Fourier Transformation (FFT) of that data set.

If considering the FD as a random process, its Hurst exponents H can be estimated and usually falls in $[1/2,2]$ (especially, $H=1/2$ when data is not with long memory). As $|k-k'|\to\infty$, the correlation between two coefficients $d_{j,k}$ and $d_{j',k'}$ decreases asymptotically as:

$$corr(d_{j,k},d_{j',k'})\sim O(\left|2^{-j}k-2^{-j'}k'\right|^{-2(p-H+1)}). \quad (4).$$

With no confusion, the coefficient $\mathbf{c},\mathbf{d}$ will be presented uniformly in the following section:

$$\mathbf{d}_i=(d_{i1},d_{i2},\cdots d_{ij}\cdots,d_{iN}), \quad (5)$$

where j is the index of wavelet basis, $\mathbf{d}_i$ is corresponding to x_i, and $N=2^J$.

Note that discrete-wavelet-based methods assume that all functions are observed at the same points, which is a normal situation. This is not a restrictive problem since we can always fit a basis and estimate the functions at the desired points.

C. Functional Data Classification and Feature extraction

Classification procedure can be split into two stages: the first stage is to abstract features for classification and the second stage is to construct classification rules. A feature vector is associated with each functional observation (FExtr stage) and this finite-dimensional vector is employed in the classification stage. Classification model is built via integrating the features and rules together.

There are two main kinds of methods of feature abstraction according to the conclusion in [11]: feature selection in which we select the best possible subset of input features and FExtr consisting in finding a transformation to a lower dimensional space [9], [12], [13]. These two methods will be combined in this paper: apply wavelet transform to the data and then select classification feature in the space transformed.

Features of data are mainly abstracted by learning in the data set. A universal aim of feature abstraction is to reduce dimension of data whereas the aim of feature abstraction for discriminant is to minimize the misdiscriminant ratio via supervised learning. Note that if ideal discriminant features are extracted (each class is represented by a region of the feature space which is well separated from the regions representative of other classes), the task of the classifier should be trivial [8]. Thus feature abstraction is a key step of classification procedure and the ability to correctly classify the test observations depend mostly on the output of the FExtr. Reference [8] discusses how to transform each observation into an appropriate vector of characteristics that represents data better. This kind of preprocessing is a powerful method for improving the performance of a learning algorithm, instead of using the raw features [14].

Wavelet based reduction is one of filtering method. Roughly, filtering reduces the infinite dimension of the observations by considering only the first coefficients of the data expanded on an appropriate wavelet basis. This approach was used by [1], [3]-[5], etc. Using wavelet based shrinkage reduction, a low dimension representation of FD can be obtained, whereas preserving as much information of data as possible, reducing to as low dimension as possible. Additionally, each component of the representation lays out the characters of data from various view point and is independent to others.

All wavelet based shrinkage methods follow these two principles: First, the reconstructed signals using fewer number of wavelet coefficients provide a very reasonable approximation to the original data. In other words, the selected wavelet coefficients are rather representative in most of the data analysis. Second, the large magnitude wavelet coefficients (in their absolute value) will characterize each signal patterns better and retain more information.

III. GA Based Feature Extraction

To extract useful features is the important way to reduce classification error and enhancing classification efficiency. Shrinkage methods represent data with low dimension whereas denoising, which is useful in reducing computing complexity of classification model. However, it has less use on the main purpose of FDC, i.e., to reduce classification error. Thereby, it is asked for a new rule of FExtr in classification problem.

It is a combination optimization problem, also a NP-hard problem, to select segment coefficients from thousands of them for minimizing the classification error. Lots of papers have shown that GA is useful to solve the combination optimization problem without prior knowledge.

A. Definition of Optimization Problem of Feature Extraction

Coefficients $\mathbf{d}_i = (d_{i1}, d_{i2}, \cdots, d_{ij}, \cdots, d_{iN})$ corresponding to X_i are obtained via transforming all FD X_i on wavelet basis funcgtions $\{\varphi_1, \varphi_2, \cdots, \varphi_{\mathrm{N}}\}$. Then FExtr procedure is executed to select fractional coefficients $\hat{\mathbf{d}}_i = (d_{is_1}, d_{is_2}, \cdots, d_{is_l}, \cdots, d_{is_L})$ for classification, where $S = \{s_1, s_2, \cdots, s_l, \cdots, s_L\} \subset \{1, 2, \cdots, N\} \triangleq S_N$.

The selected features should be comparable between different functions. Therefore, the selection of wavelet coefficients should be consistent, that is, $\forall i$, S represents the same basis positions across different functions.

Def5 : Feature extraction Vector (FEV)

Suppose that $\hat{\mathbf{d}}_i = (d_{is_1}, d_{is_2}, \cdots, d_{is_M})$ is the classification feature vector, $I = (a_1, \cdots, a_N)$ is defined as the FEV, where $a_j = 1$ if $\exists s_l \in S$ s.t. $s_l = j$, otherwise $a_j = 0$.

Obviously, there is a determinate FEV associated with each selection scheme; hence, selecting the best classification features is equivalent to find the optimal FEV.

Def6 : Optimization Object of Feature extraction

FExtr and classification is of the same object: minimizing the classification error. Therefore, the optimal object of FExtr is defined as follow:

$$f(S) = \min_{S \in S_N} [\frac{1}{m} \sum_{i=n+1}^{n+m} 1_{[g^{(S)}(X_i^{\{S\}}) \neq Y_i]}], \quad (6)$$

where $X_l^{(S)}$ denotes the classification features of X_i extracted via the FEV S and $g^{(s)}$ is the classification rule function with S under certain classifier.

Moreover, considering the number limit of features under certain condition, the object function can be attached with a penalization term $\lambda C(S)/N$ where $C(S)$ is the number of elements in S.

Def7 : Optimization Problem of FExtr (OPFE)

According to the definition above, the FExtr problem can be transformed into an optimization problem to which the GA adopted:

$$f(S) = \min_{S \in S_N} ([\frac{1}{m} \sum_{i=n+1}^{n+m} 1_{[g^{(S)}(X_i^{(S)}) \neq Y_i]}] + \frac{\lambda C(S)}{N})$$
$$S = \{s_1, s_2, \cdots, s_l, \cdots, s_L\} \subset \{1, 2, \cdots, N\} \triangleq S_N \quad (7)$$

B. GA Based Solution of Optimal Feature Extraction Vector

Considering the FEV I defined in Def5 : as independent variable's chromosome, the OPFE is transformed into an optimization problem (7) which can be solved via GA.

Step1: Confirm the Solution Space

Normally, the dimension N of coefficients obtained by wavelet transform of FD is very large. Therefore, the solution space of (7) is extremely huge. Some effective pretreatment of solution space can help to reduce searching complexity.

Note the properties of wavelet coefficient in II.B, wavelet basis $\{\varphi_1, \varphi_2, \cdots, \varphi_{\mathrm{N}}\}$ can be reordered according to Vertical Energy Method (or Seperability Method) into

$$\{\varphi_{\mathrm{k}_1}, \varphi_{\mathrm{k}_2}, \cdots, \varphi_{\mathrm{k}_{\mathrm{N}}}\}. \quad (8)$$

Meanwhile, the basis of little vertical energy $\left\| d_{vj_1} \right\|^2$ (or seperability $J_{12}(d_k)$) can be ignored. Generally, the coefficients of former H basis $\{\varphi_{\mathrm{k}_1}, \varphi_{\mathrm{k}_2}, \cdots, \varphi_{\mathrm{k}_{\mathrm{H}}}\}$ are sufficient to cover the information for classification. The value of H can be decided by repetitive experiments in which the dimension of feature increases following the order (8). Then, the space is reduced to a space with H -dimension, i.e., $S = \{s_1, s_2, \cdots, s_l, \cdots, s_L\} \subset \{k_1, k_2, \cdots, k_H\}$. Accordingly, FEV $I = (e_1, \cdots, e_H)$ is a binary valued vector of H dimension. Take I as the independent variable, the solution space is $\{0,1\}^H$.

Step2: Confirm the Original Solution Population

Since the unknown label of sample will be determined by only a few features commonly, the number L of nonzero e_j in original solution $I = (e_1, \cdots, e_H)$ is set to be a relative small value (e.g., 5,10 or 20 according to the problem). The population size is not recommended to be large, and usually 10 or 20 will be OK.

Step3: Confirm the Optimal Features

The approximate optimal solution $I^* = (e_1^*, \cdots, e_H^*)$ is approached via solving problem (7) by GA. According to Def5 : , we can get the optimal vector of features $S^* = \arg\min_{S \in S_N} f(S) = \{s_1^*, s_2^*, \cdots, s_{L'}^*\}$, where L' is the dimension of the features. Moreover, according to definitions in III.A, coefficient vector of optimal features is easy to extract as follows: $\hat{\mathbf{d}}_i^* = (d_{is_1^*}, d_{is_2^*}, \cdots, d_{is_{L'}^*})$.

Note that over fitting often arises in the optimal FExtr process, i.e., the features are selected optimally depending on the training data whereas losing the features of classification problem itself or mistaking disturbed features. This abates the efficiency of classifying new testing data as a result. There are two ways to avoid over fitting: firstly, adopt the policy of reserving multi-priority-solutions (PRMPS); secondly, increase amount of training data to reflect the character of classification problem itself.

Firstly, PRMPS means to save and refresh the best r

solutions $\{I_1,\cdots,I_r\}$ through out the genetic evolution process. The optimal FEV is defined as $I^*=I_1\vee\cdots\vee I_r$, where $\vee$ is the extract symbol of bitwise OR operation. Secondly, the learning result will be closer to the real model if training samples are increased. However, it is a remaining problem to determine amount of training examples.

Step4: Search for the Optimization in GA

The optimization search process is similar to the common optimization problem, which follows the flow as shown in .

Other than the classical GA flow, fitness is obtained by classifier, which calculated the misclassification rate via training and testing the training set according to every independent variable I. The classifier and its parameters are fixed along with the whole flow as shown in Fig. 1.

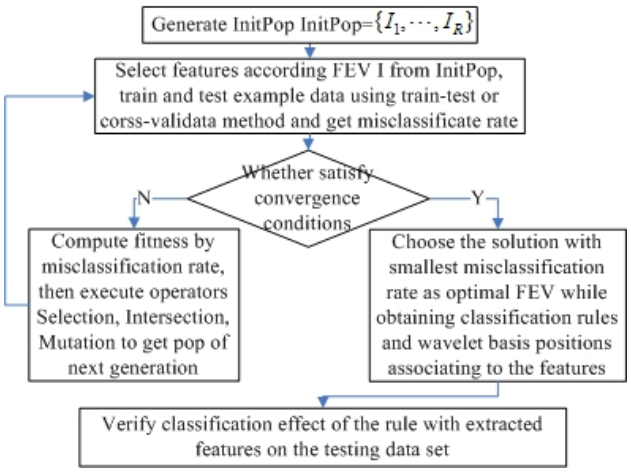

Fig. 1 Flow of Optimization Search Process

C. Convergence Analysis of Classification Error

Given rule g, the error $L_{n+m}(g)$ is expected as smaller as possible. However, it is proved by theorem 2.1 in [15] that $L_{n+m}(g)$ is larger than the Bayes probability of error L^*:

$$L^*=\inf_{g:F\to\{0,1\}}P\{g(X)\neq Y\}. \tag{9}$$

The goal of learning process is to construct rules with probability of error as close as possible to L^*. Reference [1] shows the convergence result of classification error based on vertical energy scheme:

$$E\{L_{n+m}(\hat{g})\}-L^*\le L_N^*-L^*+E\{\inf_{\substack{d=1,\cdots,N\\ g^{(d)}\in D_n^{(d)}}}L_n(g^{(d)})\}-$$
$$L_N^*+2E\{\sqrt{\frac{8\log(4S_{C_n}^N(2m))}{m}}+\frac{2}{m\log(4S_{C_n}^N(2m))}\}. \tag{10}$$

And it also has proved that $\lim_{N\to\infty}\lim_{n\to\infty}E\{L_{n+m}(\hat{g})\}=L^*$.

The same convergence result of method proposed in this paper can also be proved by similar process.

Theorem 1

Given problem with the same assumptions as Corollary 2.1 in [1], $\hat{g}^{(\hat{S})}$ is the optimal rule associated with the optimal FEV in (7) obtained from GA based training process, then rule $\hat{g}^{(\hat{S})}$ consistent for $D_n^{(S)}$ in the sense

$$\lim_{N\to\infty}\lim_{n\to\infty}E\{L_{n+m}(\hat{g}^{(\hat{S})})\}=L^*. \tag{11}$$

Proof: From the definition of wavelet transform and its decorrelation property, we know that the seperability of features is approximately additive. According to the definition of optimization problem (7), assume that the classification feature vector gained by GA is denoted by $\hat{S}$ and the classification rule function is denoted by $\hat{g}'$ when the dimension problem is not considered simultaneously in optimization process (i.e., $\lambda=0$), then we have:

$$\begin{aligned}
&E\{L_{n+m}(\hat{g}^{(\hat{S})})\}\approx E\{\min_{S\in S_N}L_{n+m}(\hat{g}^{(S)})\}\\
&=E\{\min_{S\in S_N}L_{n+m}(\arg\min_{g^{(S)}\in D_n^{(S)}}L_{n+m}(g^{(S)}))\}\\
&=E\{L_{n+m}(\arg\min_{S\in S_N,g^{(S)}\in D_n^{(S)}}L_{n+m}(g^{(S)}))\}\\
&\le E\{L_{n+m}(\arg\min_{d=1,\cdots,2^J,g^{(d)}\in D_n^{(d)}}L_{n+m}(g^{(d)}))\}\\
&=E\{L_{n+m}(\hat{g})\}
\end{aligned} \tag{12}$$

According to (10) and (12), the claim of the theorem follows via the same method of [1].

Moreover, inequation (12) accounts for stronger and faster convergence property as well as better classification effect, which own to using GA-base FExtr method. These also can be proved by experiment result analysis.

It takes longer time to training because of GA's application. However, fewer features are extracted and better effect of classification is obtained yet. As a result, it takes a little time to classify new coming observations. Commonly, the effect of classification for new examples attract more attentions in classification problem, whereas training time is not minded. So, the training time is not a balk.

IV. Experiment Analysis

To test the performance of proposed feature extraction method, we applied it to the complex classification problem (Berlinet Classification for short) in [1].

A. Process of Experiment

Step1: Randomly generate N sample data according to the definition of Berlinet Classification problem [1] with some modification.

Step2: Group example data into training set A, training-testing set C_1 and testing set C_2, which contain

examples of number N_1, N_2, and N_3 respectively.

Step3: Set parameters of GA, including initial population size R, the number L of nonzero bit in initial solution (dimension of classification features). The experiments use GA from GAOT toolbox with default operators of Selection, Intersection and Mutation.

Step4: Search for optimal solution via evolution of GA: For every solution I in population, select the features associated with I and train on set A to get the classification model. Then, test the model on set C_1 to get RCR $R(I)$ as fitness of I. Choose optimal solution $\hat{I}$ in the generation as present optimal solution. Denote the RCR of C_1 by $R_{C_1} = R(\hat{I})$. Combine A and C_1 as a general training set. Select features according to $\hat{I}$ (or I^* defined under PRMPS in III.B) and use these features to train the model of present generation. Test the model on set C_2 and get RCR R_{C_2}. The classifier K-NN (-k 3 -d 0) of MATLAB Arsenal package is used in the experiment.

Step5: Execute repetitive experiments by repeating step 1-4, then compute mean value $\overline{R}_{C_1}$, $\overline{R}_{C_2}$ of all R_{C_1}'s and R_{C_2}'s.

Step6: Use method of [1] to obtain the basis order $\{\varphi_{k_1}, \varphi_{k_2}, \cdots, \varphi_{k_N}\}$ as (8). Take coefficients of former FN ($FN = \{1, 2, \cdots\}$) wavelet basis as classification features and compute the RCRs R'_{C_1}, R'_{C_2} as well as their mean values $\overline{R}'_{C_1}$, $\overline{R}'_{C_2}$ on sets C_1, C_2 respectively.

Step7: Denote the generation time of step 4 as GT. Plot curves of $R_{C_1}, R_{C_2}, \overline{R}_{C_1}, \overline{R}_{C_2}$ along with GT and curves of $R'_{C_1}, R'_{C_2}, \overline{R}'_{C_1}, \overline{R}'_{C_2}$ along with FN.

B. *Experiment Result Analysis*

Four samples are shown in the following Fig. 2. Each curve is consisted of two different but symmetric signals, and the problem is to detect whether the two signals are close (class 1) or enough distant (class 2).

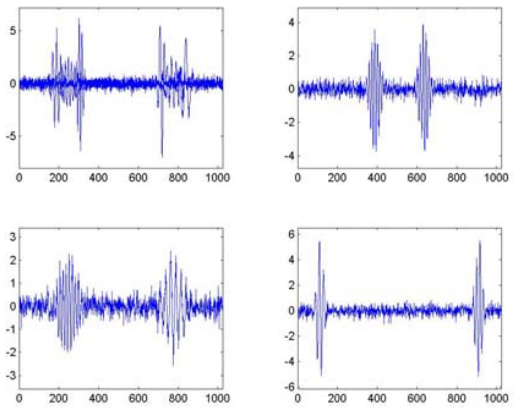

Fig. 2 Demonstration of Berlinet classification data

Using both methods proposed in [1] and in this paper respectively, the results with several parameters are shown in following Fig. 3-Fig. 6.

In Fig. 3, the abscissa is the increasing dimension of selected classification features (FN) and the vertical is the right classification rate (RCR). Dash curve line1 and dash-dot line2 show the RCR R'_{C_1}, R'_{C_2} of once experiment. Meanwhile, dot curve line3 and solid curve line4 represent $\overline{R}'_{C_1}$ and $\overline{R}'_{C_2}$, mean of RCR.

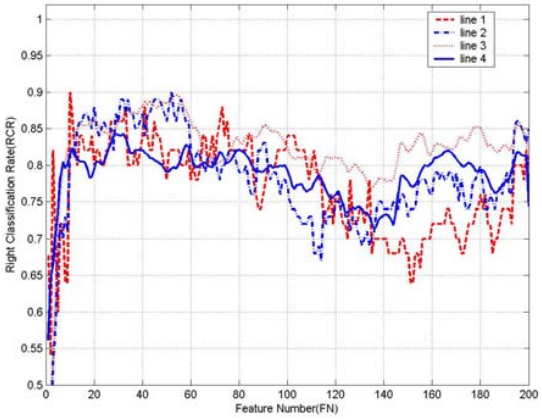

Fig. 3 Classification result of Berlinet method

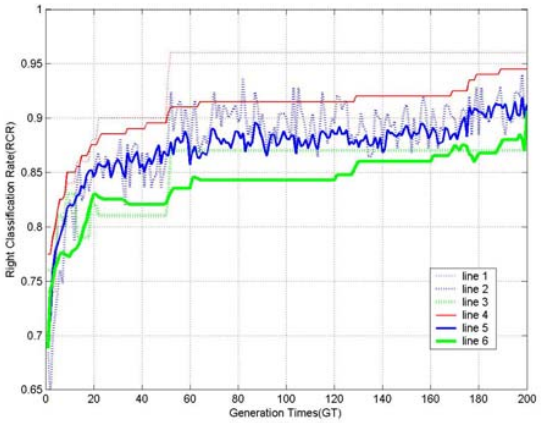

Fig. 4 Classification result of this paper

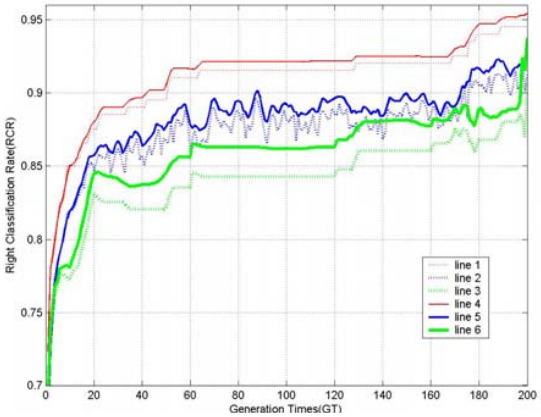

Fig. 5 Classification results using larger training set

In Fig. 4, the abscissa is the increasing generation time (GT) and the vertical is RCR. Dot curve line1, line2, line3 are corresponding to R_{C_1}, R_C, R_{C_2} respectively and solid curve line4, line5, line6 are corresponding to $\overline{R}_{C_1}$, $\overline{R}_C$, $\overline{R}_{C_2}$, mean RCR of present generation, where R_C is mean of all R_{C_1} in population of present generation and $\overline{R}_C$ is the mean of R_C's.

In Fig. 5, the abscissa is the increasing generation time (GT) and the vertical is RCR. Solid curves line4, line5, and line6 show the results $\overline{R}_{C_1}$, $\overline{R}_C$, and $\overline{R}_{C_2}$ when using larger training set. And dot curve line1, line2, line3 represent results $\overline{R}_{C_1}$, $\overline{R}_C$, $\overline{R}_{C_2}$ as same as in Fig. 4.

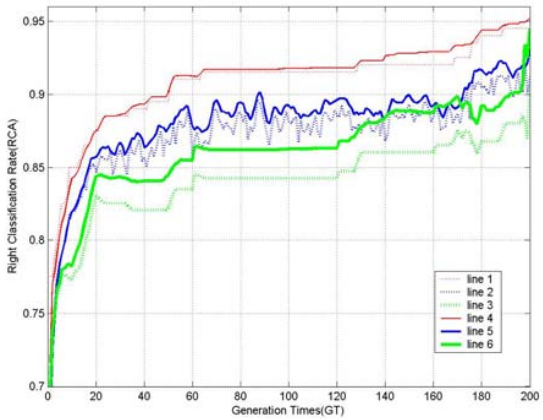

Fig. 6 Classification result using PRMPS

The solid curve in Fig. 6 show results of PRMPS. Other curves represent the same results as Fig. 5.

According to line1 of Fig. 4, the method of this paper makes R_{C_1} R_{C_2} converge to 0.96 and 0.875 rapidly. Compared with Fig. 3 showing result of method from [1], the use of GA can extract optimal classification features faster and the efficiency of classification can be enhanced obviously when applying GA based FExtr method. According to the variance of R_C, mean value of population, it is easy to find that RCR corresponding to FEV tends to approach to the optimum during evolution. Mean values $\overline{R}_{C_1}$, $\overline{R}_C$, $\overline{R}_{C_2}$ gained from repetitive experiments are also steady which is the reflection of effect of our method. On the other hand, R_{C_2} of C_2 is higher compared with R'_{C_2} whereas relatively lower than R_{C_1}. This is the evidence that over fitting problem exists in optimization FExtr Fig. 5 and Fig. 6 show the results using larger training set and PRMPS. These figures suggest that the gap between R_{C_2} and R_{C_1} gained by former methods shrinks relatively. The over fitting problem is solved in some sense. Over fitting is an inherent difficult problem of learning algorithm. It is a remaining problem that no method can solve completely.

Appendix

Definition of Berlinet Classification Problem

For each $i=1,\cdots,n$, the functional data and their class labels $(X_i(t),Y_i)$ are generated via the following scheme:

$$X_i(t)=\tfrac{1}{50}(\sin(F_i^1 t)\,f_{\mu_i,\sigma_i}(t)+\sin(F_i^2 t)\,f_{\mu_i',\sigma_i}(t))+\varepsilon_i$$

where $f_{\mu,\sigma}$ stands for the normal density with mean μ and variance σ^2; F_i^1 and F_i^2 are uniform random variables on $[50,150]$; μ_i and σ_i are randomly uniform respectively on $[0.1,0.4]$ and $[0,0.005]$; $\mu_i'=1-\mu_i$; and the ε_i's are mutually independent normal random variables with mean 0 and standard deviation 0.5. The label Y_i associated to X_i is then defined to be $Y_i=0$ when $\mu_i\le 0.25$ and $Y_i=1$ otherwise.

References

[1] A. Berlinet, G. Biau, and L. Rouvière, "Functional supervised classification with wavelets," *Annales de l'ISUP*, vol. 52, 2008, pp. 61-80.
[2] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis.* Springer, New York, 2005
[3] P. N. Belhumeur, J. P. Hepana, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis, and Machine Intelligence*, vol.19 1997, pp.711–720.
[4] P. Hall, D. S. Poskitt, and B. Presnell. "A functional data-analytic approach to signal discrimination," *Technometrics*, vol. 43, 2001, pp.1–9.
[5] U. Amato, A. Antoniadis, and I. D. Feis, "Dimension reduction in functional regression with applications," *Computational Statistics and Data Analysis*, vol. 50, 2006, pp. 2422–2446.
[6] F. Ferraty and P. Vieu, *Nonparameter Functional Data Analysis: Theory and Practice*, Springer, 2006.
[7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Springer, New York, 1997.
[8] Irene Epifanio, "Shape descriptors for classification of functional data," *Technometric*, vol. 50, no. 3. 2008.
[9] G. Rosner and B. Vidakovic, "Wavelet functional ANOVA, Bayesian false discovery rate, and longitudinal measurements of Oxygen," Pressure in Rats, Technical Report 1/2000, ISyE, Georgia Institute of Technology, 2000
[10] S.G. Mallat, *A Wavelet Tour of Signal Processing*, San Diego: Academic Press, 1998.
[11] Marek Kurzynski and Edward Puchala, "The optimal feature extraction procedure for statistical pattern recognition," *ICCSA 2006, LNCS 3982*, pp. 1210–1215.
[12] C. Abraham, G. Biau, and B. Cadre, "On the kernel rule for function classification," *Annals of the Institute of Statistical Mathematics*, vol. 58, 2006, pp. 619–633.
[13] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, 2005, pp.323–375.
[14] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," *Data mining, inference and prediction*, Springer-Verlag, 2001
[15] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New-York, 1996.