International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:4, No:1, 2010

Wavelet and K-L Seperability Based Feature Extraction Method for Functional Data Classification

Jun Wan, Zehua Chen, Yingwu Chen, Zhidong Bai

***Abstract*—**This paper proposes a novel feature extraction method, based on Discrete Wavelet Transform (DWT) and K-L Seperability (KLS), for the classification of Functional Data (FD). This method combines the decorrelation and reduction property of DWT and the additive independence property of KLS, which is helpful to extraction classification features of FD. It is an advanced approach of the popular wavelet based shrinkage method for functional data reduction and classification. A theory analysis is given in the paper to prove the consistent convergence property, and a simulation study is also done to compare the proposed method with the former shrinkage ones. The experiment results show that this method has advantages in improving classification efficiency, precision and robustness.

***Keywords*—**classification, functional data, feature extraction, K-L seperability, wavelet.

I. Introduction

FUNCTIONAL data analysis (FDA) is a new developed branch of statistics which attracts more and more attention in many fields, such as industry control, information management, simulation experiment, etc., because more and more data generated in these field are often in the form of long time series, continuous factors depending. Moreover, many successful cases of FDA have also suggested its advantage.

Functional data classification (FDC) is an interesting problem in FDA, because many science and application problems, such as recognition, prediction, control, decision making, management, etc., end up with classification problems. In many real-life problems, input data are in fact (sampled) functions rather than standard d-dimensional vectors, and this casts the classification problem into the class of FDA [1]. Classification is one of two common goals in application of FDA [2].

It is a key problem of FDA, including FDC, to reduce dimension and correlation of functional data (FD) simultaneously whereas keeping its functional features, such as integrality and smoothness. A standard answer to both problems of FD is to extend PCA [3] or ICA [4] method as well as to extend wavelet methods [1], [5]. More and more studies show that wavelet-based methods, namely discrete wavelet transform (DWT) and shrinkage methods, are suitable to solve the problem above as the nice properties of wavelet: smoothness, multi-scale time-frequency decomposition, orthogonality, vanishing moments [1], [6]-[8], etc.

A universal aim of feature extraction is to reduce dimension of data. Shrinkage method presents good performance to keep global characters and denoise in low-dimension FD representation. However, the aim of feature abstraction for discriminant is to minimize the misdiscriminant ratio via supervised learning, which is not concerned in the shrinkage method. Shrinkage method gives a low dimension representation of FD which contains most information of data, whereas not all information is needed for classification in fact. Moreover, shrinkage method prefers reserving features with large power for reconstructing the function, whereas there is still the possibility that some of the discarded features with small power may be non-trivial discriminatory.

Consequently, to extract 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. The extraction method is required to abstract and select features with large power (to keep functional information) and large discrimination (to keep classification information). According to the information theory, discrimination of feature can be measured by some specially defined seperability upon the training data.

In this paper, a novel method using K-L seperability order to extract classification features and reduce the dimension is proposed. This method is an advanced approach of the popular wavelet based shrinkage method for functional data reduction. It is proved by theory analysis and simulation experiment that this method has advantages in improving classification efficiency, precision and robustness.

II. Problem Definition and Background

A. Basic Definition and Hypothesis

The problem of classification is about guessing or predicting the unknown class of an observation. An observation is a collection of measurements represented by functional data in the field of FDA.

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).

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 [9]. 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 $F = L_2([0,1])$ (that is, 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. 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 [10]: 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 [5], [11], [12]. 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 [4]. 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 [4] discussed 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 [13].

III. Wavelet Based Functional Representation via Low Dimension Features

A. 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 $\{\phi_v\}$.

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 [3] or ICA [4] method as well as to extend wavelet methods [1], [5]. 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 in this paper.

B. Wavelet-Based Functional Representation

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 [14]. 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 $\{0, \pm1, \pm2, \cdots\}$, 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:

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

where $J > L$ and L is 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 $\mathbf{c}_L = (c_{L,0}, \cdots, c_{L,2^L-1})$, $\mathbf{d}_L = (d_{L,0}, \cdots, d_{L,2^L-1})$, $\mathbf{d}_J = (d_{J,0}, \cdots, d_{J,2^J-1})$ 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. Wavelet Coefficient Shrinkage and Low Dimension Representation

Wavelet based reduction is one of filtering methods. Roughly, filtering reduces the infinite dimension of the observations by considering only the first d coefficients of the data expanded on an appropriate wavelet basis. This approach was used by [1], [6]-[8], 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.

In order to selecting wavelet coefficients for different single curves heterogeneously, reference [15] introduced a Wavelet Vertical Energy metric of multiple curves and utilized it for the efficient data reduction as well as the following FDA problem, which proposed the following data-reduction criteria with goals of minimizing Overall Relative Reconstruction Error (ORRE):

$$ORRE(\lambda) = \frac{\sum_{j=1}^{N} E[\left\| d_{vj}(1 - I(\left\| d_{vj} \right\|^2 > \lambda)) \right\|^2]}{\sum_{j=1}^{N} E[\left\| d_{vj} \right\|^2]} + \xi \cdot \frac{\sum_{j=1}^{N} E[\left\| I(\left\| d_{vj} \right\|^2 > \lambda)) \right\|]}{m}, \quad (6)$$

where λ is the threshold parameter and $\left\| d_{vj} \right\|^2$ is the sum of

all wavelet coefficients at the j -th wavelet-position:

$$\left\|d_{vj}\right\|^2 = \sum_{k=1}^{m} d_{kj}^{\ 2} . \quad (7)$$

The wavelet-positions of vertical energy larger than λ are selected according to $I(\left\|d_{vj}\right\|^2 > \lambda)$ in (6).

IV. Classification Feature Extraction Based on K-L Seperability

The extraction of features is the best way to reduce classification error and enhancing classification efficiency, which has important influence on classification. 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 FD classification, i.e., to reduce classification error. Thereby, it is asked for a new rule of FExtr in classification problem.

A. Definition and Properties of K-L Seperability

Seperability is one way to evaluate the classifying ability of some features. Seperability is commonly defined as some distance or dissimilarity between two classes.

Supposing that functions belong to two classes ω_1, ω_2, the feature of function, denoted by d, belongs to the two classes with probability $P_1(d) = P(d \mid \omega_1)$, $P_2(d) = P(d \mid \omega_2)$. According to the information theory [16], the information entropy of class c is denoted by $H_c(x) = \int P_c(x) \ln P_c(x) dx$, supposing the probability density function of c is $P_c(x)$. K-L distance is defined as the relative entropy of probability density function, i.e.,

$$D_{12} = D(P_1 \| P_2) = \int P_1(x) \ln \frac{P_1(x)}{P_2(x)} dx .$$

This distance is the entropy of density function P_2 relative to P_1. Similarly, the relative entropy of P_1 to P_2 can also be obtained. Entropy D_{12} presents the dissimilarity of probability density function. In other words, D_{12} is one measurement of probability difference that a feature belongs to two classes. Therefore, it can be treated as an standard of one feature to separate two classes.

Def5 K-L Seperability

To be symmetric, the K-L seperability of feature x on classes ω_1, ω_2 is defined as sum of the relative entropy of P_1, P_2:

$$J_{12} = D_{12} + D_{21} = \int (P_1(x) - P_2(x)) \ln \frac{P_1(x)}{P_2(x)} dx \quad (8)$$

Property1:

K-L seperability is of the nonnegative property, symmetry property, and additive independence property, i.e.,

a) $J_{12} \geq 0$; b) $J_{12} = J_{21}$; c) $d_1, d_2, \cdots, d_m$ $i.i.d \Rightarrow$ $J_{12}(d_1, d_2, \cdots, d_m) = \sum_k J_{12}(d_k)$.

Property2:

The coefficients with larger K-L seperability contain more classification information.

Property3:

Suppose that the two classes are characterized separately by n-dimension vector $\mathbf{d}$ which distributed as n-dimension normal $N(\theta_1, \sum_1)$, $N(\theta_2, \sum_2)$, where θ_1, θ_2 are mean vectors and $\sum_1, \sum_2$ are covariance matrixes. Then, K-L seperability can be rewritten as follows:

$$\begin{aligned} J_{12}(\mathbf{d}) &= \frac{1}{2} tr[(\sum_1 - \sum_2)(\sum_1^{-1} - \sum_2^{-1})] \\ &+ \frac{1}{2} tr[(\sum_1^{-1} - \sum_2^{-1})(\theta_1 - \theta_2)(\theta_1 - \theta_2)'] \end{aligned} \quad (9)$$

and $j_{ij}(\mathbf{d}) = (\theta_1 - \theta_2)' \sum^{-1} (\theta_1 - \theta_2)$ as $\sum_1 = \sum_2 = \sum$. Specially, K-L seperability is easy to obtain when $\mathbf{d}$ is only one dimension, i.e.,

$$J_{12}(d) = \frac{(\theta_1 - \theta_2)^2}{\sigma^2} \quad (10)$$

where θ_1, θ_2 are means and σ^2 is the variance.

B. K-L Seperability Based Wavelet Basis Selection

Wavelet basis selection is to choose the coefficients of functions as the classification characters, which should be comparable between different functions. Therefore, the selection of wavelet coefficients should be consistent, that is, to select coefficients of the same basis positions across different functions.

In this paper, the selection rule of wavelet basis is based on vertical energy order and K-L seperability order. The wavelet coefficient positions, in terms of vertical energy defined as (5) and (7), are ranked as follow:

$$\left\|d_{vj_1}\right\|^2 \geq \left\|d_{vj_2}\right\|^2 \geq \cdots \geq \left\|d_{vj_N}\right\|^2 \quad (11)$$

Meanwhile, in terms of K-L seperability defined as (10), the positions are reordered as:

$$J_{12}(d_{j_1'}) \geq J_{12}(d_{j_2'}) \geq \cdots \geq J_{12}(d_{j_N'}) \quad (12)$$

The order indexes $p_1, p_2, \cdots, p_N$ and $p_1', p_2', \cdots, p_N'$ of wavelet basis positions are obtained via scheme (11) and (12) where $p_s = k \ if \ j_k = s$, $p_s' = k \ if \ j_k' = s$.

The basis functions $\{\varphi_1, \varphi_2, \cdots, \varphi_N\}$ are reordered into $\{\varphi_{k_1}, \varphi_{k_2}, \cdots, \varphi_{k_N}\}$ by combining (11) and (12) so that

$$\begin{aligned} &\lambda p_{k_1} + (1-\lambda) p_{k_1}' \leq \lambda p_{k_2} + (1-\lambda) p_{k_2}' \\ &\leq \cdots \leq \lambda p_{k_N} + (1-\lambda) p_{k_N}' \end{aligned} \quad (13)$$

Then, the coefficients $\mathbf{d}_i$ obtained from wavelet transform of functional data X_i can be ranked according above order, rewriting as $X_i' = (X_{ik_1}, X_{ik_2}, \cdots, X_{ik_N})$.

Given a classification method and an arbitrary dimension d, $X_i^{(d)} = (X_{ik_1}, X_{ik_2}, \cdots, X_{ik_d})$ can be used as the characters of X_i and the classification rules can be denoted as $D_n^{(d)}$:

$$D_n^{(d)} = \{g^{(d)} : R^d \times (R^d \times \{0,1\})^n \to \{0,1\}\}.$$

According to the scheme of (13), the selection of classification characters depends on λ and d. It is the vertical energy method [1] as $\lambda = 0$. When $\lambda = 1$, the characters are selected only based on K-L seperability proposed in this paper. To give attentions both on fitting precision and seperability of classification characters, λ can take value $1/2$. Cross validation method can be used to choose λ in practice. Feature dimension d and corresponding rule $g^{(d)}$ are approached via scheme as follow:

$$(\hat{d}, \hat{g}^{(\hat{d})}) \in \arg\min_{\substack{d=1,\cdots,N \\ g^{(d)} \in D_n^{(d)}}} [\frac{1}{m} \sum_{i=n+1}^{n+m} 1_{[g^{(d)}(X_i^{(d)}) \neq Y_i]}]. \quad (14)$$

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 [17] that $L_{n+m}(g)$ is larger than the Bayes probability of error L^*:

$$L^* = \inf_{g:F \to \{0,1\}} P\{g(X) \neq Y\}. \quad (15)$$

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 (i.e., $\lambda = 0$ in our method):

$$E\{L_{n+m}(\hat{g})\} - L^* \leq L_N^* - L^* + E\{\inf_{\substack{d=1,\cdots,N \\ g^{(d)} \in D_n^{(d)}}} L_n(g^{(d)})\}$$
$$-L_N^* + 2E\{\sqrt{\frac{8\log(4S_{C_n}^N(2m))}{m}} + \frac{2}{m\log(4S_{C_n}^N(2m))}\} \quad (16)$$

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], $\hat{g}'$ is the optimal rule defined in (14) obtained from training process, then $\hat{g}'$ consistent for $D_n^{(d)}$ in the sense

$$\lim_{N\to\infty} \lim_{n\to\infty} E\{L_{n+m}(\hat{g}')\} = L^*. \quad (17)$$

Proof: From the definition of K-L seperability (8) and its additive independence *property1*, we know that larger seperability accounting for smaller classification error under corresponding characters. According to the order scheme of (13), $\forall d$, we have

$$E\{L_{n+m}(\hat{g}')\} \leq E\{L_{n+m}(\hat{g})\}. \quad (18)$$

According to (5) and (7), the claim of the theorem follows via the same method of [1].

Moreover, inequation (18) accounts for stronger and faster convergence property as well as better classification effect which own to using our character extraction method. These also can be proved by experiment result analysis.

V. Experiment Analysis

To test the performance of proposed feature extraction method, we applied it to Mallat piecewise functions classification problem [14] as well as the complex classification problem (Berlinet Classification for short) in [1].

According to the definition of Mallat piecewise function and Berlinet classification data, 100 sample data were generated for each group. Four samples from each group are shown in the following Fig.1 and Fig.2.

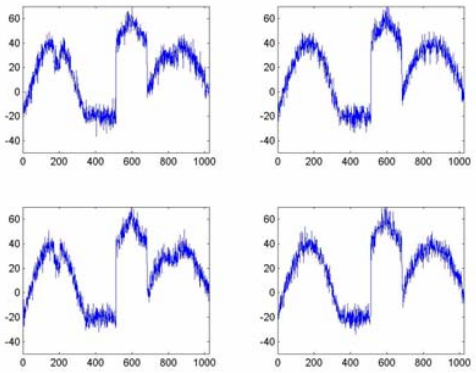

Fig. 1 Demonstration of Mallat piecewise function data

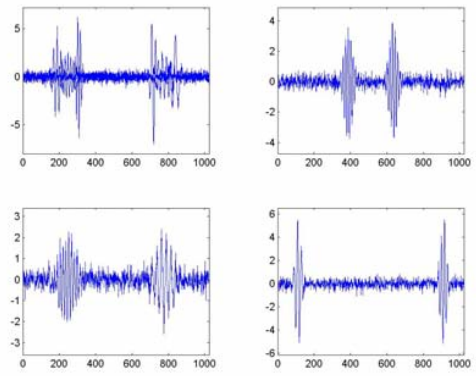

Fig. 2 Demonstration of Berlinet classification data

The first group of data were generated using two classes of Mallat piecewise function with additional error on each point which is randomly uniform in [-5,5]; and the second group

were generated on definition of [1] with some modification.

The difference of two classes in the first group is the depressed parts around abscissa 200 and 800 in one class. As shown in Fig.2, each curve of the second group is composed of two different but symmetric signals, and the problem is thus to detect if the two signals are close (class 1) or enough distant (class 2). The second group is more complex than the first one.

In the experiment, each group of data are split into training set and testing set, and each set contains 50 samples. Afterwards, let the training samples decide the classification characters and rule by itself and apply the result on testing set to get the testing error. In above process, the classifier is chosen to be K-NN (kNN_classify -k 3 -d 0) of MATLAB Arsenal.

Using method proposed in this paper and setting $\lambda = 0,1/2,1$ separately, the results of classification experiment for above-mentioned data are shown in Fig.3, Fig.4 and Fig.5. Picture of $\lambda = 0$ was obtained using vertical energy method of [1] and pictures of $\lambda = 1/2,1$ were obtained using our new method (two special cases). In all these figures, the abscissa is the dimension of selected classification characters (FN) and the vertical is the right classification rate (RCR).

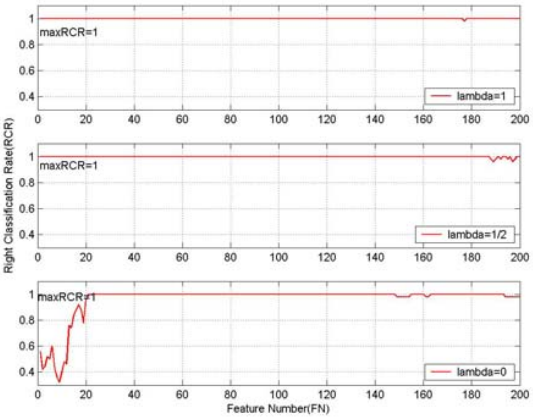

Fig. 3 Classification results of Mallat piecewise functions

The classification result of Mallat piecewise function data is shown in Fig.3, from which we can find easily that excellent RCR can be obtained by using only a little classification features when $\lambda = 1/2,1$. Mass experiments showed that the method proposed in this paper can extract the features which represent the difference between classes efficiently. Moreover, these experiments also showed that only a little classification features will count for a great deal in FDC, when the classification problem is similar to Mallat piecewise functions, i.e., there are only some local differences between classes.

The experiment results of Berlinet classification problem are shown in Fig.4 and Fig.5, which used Train-Test-Validate method and cross-validate (3 times) respectively. In these figures, solid curve "mean" estimated mean of RCR for 50 replications of the experiments, and dash curve "once" is one result of them. Since the complexity of Berlinet classification problem, it is difficult to extract useful classification features. From the figures, it is shown that the results obtained as $\lambda = 1/2,1$ are better than those obtained as $\lambda = 0$, which contrast is more obvious in Fig.4 and Fig.5. Moreover, from the trend of RCR varying with FN, we can conclude that better classification efficiency can be approached faster and more steadily as $\lambda = 1/2,1$ than those as $\lambda = 0$. These are evidences that the feature extraction method proposed in this paper can help to boost up classification robustness whereas to accelerate the classification process.

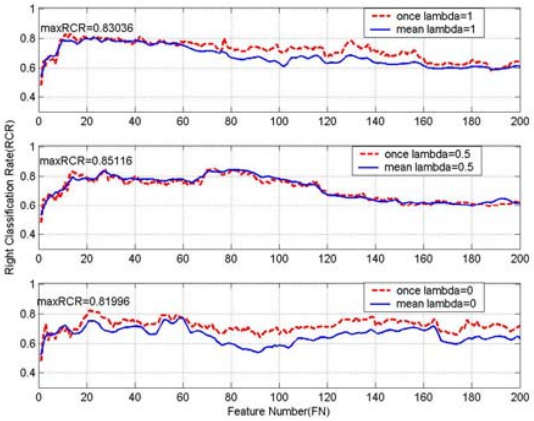

Fig. 4 Berline Classification Result (Train_Test_Validate)

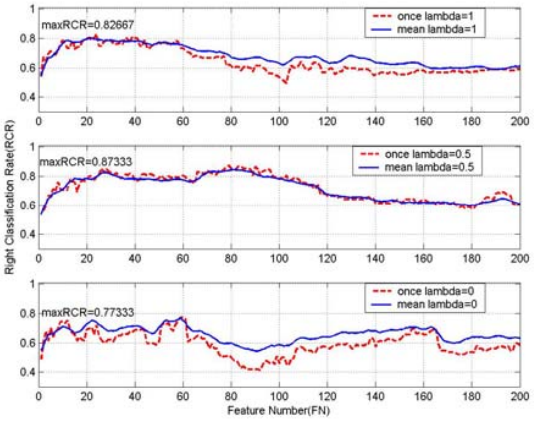

Fig. 5 Berline Classification Result (cross_validate 3 times)

APPENDIX

A. *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.

B. Definition of Mallat Piecewise Function 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:

$X_i(t) = X^*(t) + \varepsilon_i$ as $Y_i = 0$;

$$X_i(t) = \begin{cases} X^*(t) - 15 + \varepsilon_i & (1/6 < t \le 1/5) \\ 28 & (3/4 < x \le 5/6) \text{ as } Y_i = 1; \\ X^*(t) & else \end{cases}$$

where

$$X^*(t) = \begin{cases} 60\sin(3\pi t) - 20 & (0 \le t \le 1/3) \\ -20 & (1/3 < t \le 1/2) \\ 20 \times 2^{4t} - 40 & (1/2 < t \le 7/12) \\ 20 \times 2^{14/3 - 4t} - 40 & (7/12 < t \le 2/3) \\ -360(2t - 5/3)^2 + 38 & (2/3 < t \le 1) \end{cases}$$

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] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Springer, New York, 1997.

[4] Irene Epifanio, "Shape Descriptors for Classification of Functional Data," *Technometric*, vol. 50, no. 3. 2008.

[5] 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

[6] 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.

[7] P. Hall, D. S. Poskitt, and B. Presnell. "A functional data-analytic approach to signal discrimination," *Technometrics*, vol. 43, 2001, pp.1–9.

[8] 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.

[9] F. Ferraty and P. Vieu, *Nonparameter Functional Data Analysis: Theory and Practice*, Springer, 2006.

[10] Marek Kurzynski and Edward Puchala, "The optimal feature extraction procedure for statistical pattern recognition," *ICCSA 2006, LNCS 3982*, pp. 1210–1215.

[11] 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.

[12] 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.

[13] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," *Data mining, inference and prediction*, Springer-Verlag, 2001

[14] S. G. Mallat, *A Wavelet Tour of Signal Processing*, San Diego: Academic Press, 1998.

[15] U. K. Jung, M. K. Jeong, , J.C. Lu, "Wavelet-based Data Reduction and Mining for Multiple Functional Data," *International Journal of Production Research*, vol. 44, no. 14, 2006, pp. 2695-2710(16).

[16] C. R. Shalizi, "Methods and techniques of complex systems science: An overview," T. S. Deisboeck and J. Y. Kresh, *Complex Systems Science in Biomedicine*, Chapter 1, pp. 33-114, Springer, Singapore, 2006.

[17] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New-York, 1996.