Journal of Electronics and Information Science (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2023.080613
ISSN 2371-9524 Vol. 8 Num. 6

A Functional Data Classification Model Utilizing Functional Mahalanobis Distance and Regenerative Kernel Methods

Xinyu Huang[1], Ziyang Pan[2]

[1]*School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, 066004, China*
[2]*School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China*

Keywords: Functional Data Classification, Functional Mahalanobis distance, KPCA

Abstract: The classification of functional data is an important research direction in modern data mining. In this paper, we propose a similarity measurement method for functional data based on functional Mahalanobis distance and regenerative kernel theory, considering the scenario where the predictor variable is a random function and the response variable is a categorical scalar. This method is then applied to functional kernel principal component analysis. During the classification phase, classic algorithms such as support vector machines and random forests can be combined to accomplish the task of classifying functional data. In empirical analysis, compared to the regenerative kernel based on Euclidean distance and the Euclidean distance regenerative kernel based on B-spline basis functions, the proposed method achieves better classification results. Furthermore, this similarity measurement can also be utilized in other machine learning algorithms based on regenerative kernel theory, thereby developing corresponding analysis methods for functional data.

1. Introduction

With the continuous development of human society and the Internet, data collection techniques have made significant advancements in the past few decades, enabling the collection of continuous dynamic data. [1~2] In 1982, Canadian statistician Ramsay introduced the concept of functional data,[3] which refers to data where the observed values exhibit linear or nonlinear, as well as multidimensional relationships manifested as smooth curves or continuous functions. Compared to traditional multivariate data, functional data have the characteristic of infinite dimensionality.

The classification of functional data methods is essentially an extension of multivariate statistical methods and can generally be divided into filtering and regularization. Filtering involves selecting a suitable set of basis functions, and then using traditional classification methods for the coefficients. For example, the function data can be expanded using B-spline basis functions, and then the coefficients of the basis functions can be classified using support vector machine (SVM) algorithm. The final effectiveness of this method depends entirely on the choice of basis functions. For instance,

Pourshoghi et al[4]. used B-spline basis functions to expand the data and then applied SVM for classification. Regularization involves feature selection on functional data to obtain the most crucial data points or intervals, achieving dimension reduction from infinite to finite dimensions, followed by using machine learning algorithms for classification. For example, Jin Haibo and Ma Haiqiang[5] proposed a segmented method for extracting feature points from functional data; Pini and Vantini[6] presented a hypothesis testing-based method for interval classification of functional data. In addition to these two main strategies, many researchers have explored the classification of functional data from other perspectives. For example Fan et al.[7] proposed an algorithm for functional data classification based on the random forest algorithm; Rossi and Villa[8] combined the kernel method with the support vector machine algorithm to classify functional data; Thind et al.[9] based on feed-forward neural network to classify functional data; Fuchs et al.[10] Classification of functional data based on nearest neighbor classification method; Rossi et al.[11] Classification of functional data based on multilayer perceptron machine algorithm; Ke Chien-Kun's[12] classification model for Riemannian manifold functional data; Vommi Amukta Malyada et al.[13] Fuzzy KNN Hybrid Filtering and Encapsulated Feature Selection Classification based on Bonferroni Mean.

The aforementioned methods for functional data classification are all based on specific classification techniques, which have their limitations in terms of applicability. Therefore, this study proposes a functional data similarity measure based on functional Mahalanobis distance and reproducing kernel theory, and applies this measure to kernel principal component analysis, thus projecting infinite-dimensional functions into finite-dimensional spaces. The Mahalanobis distance constructed based on the reproducing kernel effectively captures the information of functional data, depicting the differences between different functions. This similarity measure is suitable for functional data analysis methods based on reproducing kernel theory. Finally, this study verifies the effectiveness of the proposed method by applying it to practical data tasks.

2. Theory and methodology

2.1 FPCA-based Functionalized Mahalanobis Distance

In 1936 Indian statistician P.C. Mahalanobis proposed the Mahalanobis distance, which takes into account scale-independent links between various characteristics and is a measure of distance that can be regarded as a modification of the Euclidean distance for solving the covariance distance. For a multivariate vector $x = (x_1, x_2 \ldots, x_n)^T$ with mean $\mu = (\mu_1, \mu_2 \ldots, \mu_n)^T$ and covariance matrix$\sum$, its Mahalanobis distance is:

$$D(\mathrm{x}) = \sqrt{(x-\mu)^T \sum^{-1} (x-\mu)} \tag{1}$$

For random variables X and Y that are uniformly distributed and whose covariance matrix is$\sum$, the Mahalanobis distance between data points x,y is:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \sum^{-1} (\vec{x}-\vec{y})} \tag{2}$$

Next, we introduce functional principal component analysis. It follows from Mercer's theorem that there exists a continuous sequence of functions $\{v_k(t), k \geq 1\}$ and a monotonically decreasing sequence of positive numbers such that the following equation holds:

$$V(s,t) = \sum_{k=1}^{\infty} \lambda_k v_k(s) v_k(t) \tag{3}$$

For a random function $X(t)$ there is:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \varphi_k(t) \tag{4}$$

Where $\mu(t)$ is the mean function, ξ_{ik} is the centrality function, the eigenfunction $\varphi_k(t)$,k=1,2,…,∞ are a set of pairwise orthogonal basis functions in the L^2 space, and$X_i(t) - \mu(t)$represents the projection scores of the feature function $\varphi_k(t)$, i.e.:

$$\xi_{ik} = \int_0^1 (X_i(t) - \mu(t)) \varphi_k(t) dt \tag{5}$$

Satisfy $E(\xi_{ik}) = 0, D(\xi_{ik}) = \lambda_k$. In practical problems, the first K principal functions that can represent most of the information of the data are usually chosen. According to Ramasy and Silverman's method of solving for principal component scores can be obtained:

$$N^{-1} W^{1/2} C^T C W^{1/2} u = \rho u \tag{6}$$

Where ρ is the eigenvalue and u is the eigenfunction. Finally, we introduce the functional type principal component analysis based on the functional type martens distance. Firstly, the data can be centered to get $\mu(t) = 0$ and then the sample principal component function is calculated according to the method (6), and then the principal component score can be obtained by projecting the data:

$$A = \int X(t) v(t) dt \tag{7}$$

A is a matrix of N*K, which is then obtained by calculating the Mahalanobis distance for the score A:

$$D(X_i(t), X_j(t)) = \sqrt{\frac{\sum_{k=1}^{K} \{\int [X_i(t) - X_j(t)] \hat{v}_k(t) dt\}^2}{\hat{\lambda}_k}} \tag{8}$$

Where $\hat{\lambda}_k$ and $\hat{v}_k(t)$ are estimates of the variance and function of the principal component scores.

2.2 Classification models based on functional kernel principal component analysis

Suppose the original data is $X = [x_1, x_2, \ldots, x_n] \in R^{m*n}$, m is the amount of data and n is the number of variables. Firstly, the characteristic covariance matrix can be obtained by mapping as:

$$C = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \phi(x_i)^T \tag{9}$$

Satisfies: $\lambda V = CV$ .Where λ is the eigenvalue and V is the eigenvector. Since C is unknown, the eigenvalues and eigenvectors cannot be derived. It can be solved by introducing a nonlinear transformation, i.e.:

$$\lambda \phi(x_j) V = \phi(x_j) CV \tag{10}$$

And satisfy the existence of $\alpha_i (i = 1,2, \ldots, n)$ such that the following equation holds:

$$V = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \phi(X) \alpha \tag{11}$$

Where $\phi(X) = [\phi(x_1), \phi(x_2), \ldots, \phi(x_n)]; \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T$, and thus Eq. (11) can be written

as:

$$\lambda\phi(X)\alpha = \frac{1}{n}\phi(X)\phi(X)^T\phi(X)\alpha \tag{12}$$

The left-multiplication of both ends by $\phi(X)^T$ gives:

$$\lambda\phi(X)^T\phi(X)\alpha = \frac{1}{n}\phi(X)^T\phi(X)\phi(X)^T\phi(X)\alpha \tag{13}$$

Bringing in the kernel function $K = K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$,yields: $\lambda\alpha = K\alpha$.The kernel function adopted in this paper is the functional martensitic distance, i.e:

$$K(x_i, x_j) = \sqrt{\frac{\sum_{k=1}^{K}\{\int[X_i(t) - X_j(t)]\hat{v}_k(t)dt\}^2}{\hat{\lambda}_k}} \tag{14}$$

3. Data analysis

3.1 Background of empirical data

ECG is a medical test used to diagnose heart diseases and abnormalities by recording the heart's electrical signals in a graphic manner. These electrical signals are generated by the electrical activity of the heart muscle.ECG visualizes this activity as a series of waveforms, usually including P waves, QRS waves, and T waves.ECG is widely used in clinical medicine.The shape, duration, and intervals of these waveforms provide important information about the health of the heart. Its main uses include: detecting abnormalities in the heart's rhythm; helping doctors determine whether a patient's heart rhythm is normal; showing signs of myocardial infarction, such as ST-segment elevation or lowering; for long-term monitoring of cardiac function; doctors can use ECGs to assess the effects of specific medications or therapeutic measures on a patient's cardiac function; and before a surgery, a doctor may order ECGs to ensure that the patient's cardiac health is suitable for surgery, etc.

ECG generates a huge amount of data, including hundreds of heartbeat waveforms, which need to be classified in order to better understand and utilize the data. By categorizing ECG data, doctors can identify heart diseases and abnormalities in order to make an accurate diagnosis, with different types of heart problems requiring different treatments; researchers use categorized ECG data to study heart health and new treatments for heart disease; and regular ECG monitoring of patients can help doctors detect potential heart problems early and take preventive measures.

3.2 Data sources

The data comes from http://timeseriesclassification.com/description.php?Dataset=ECG200, a dataset provided by R. Olszewsk from 2001, where each class tracks the electrical activity recorded during one heartbeat. As a binary classification model, these two classes are are normal heartbeat and myocardial infarction, and both the training and test sets are 100 data and 96 in length.

3.3 Comparison of experimental setup and results

In this paper, a binary actual functional dataset is used to test the algorithm of this paper, the experiments compare the regenerative kernel model of this paper based on functional martens distance with the algorithms of regenerative kernel model based on B-spline Euclidean distance and regenerative kernel model based on Euclidean distance and the results of the comparisons are

expressed in terms of accuracy. All the experiments in this paper are realized using R language programming, and the experiments will be repeated 100 times, and then the average as well as the variance of the results of the 100 experiments will be taken to get the final experimental results. The experimental data is divided into training set and test set using 1:1 ratio and then experiments are conducted.

Table 1: Confusion Matrix

		Reference	
		Positive	Negative
Prediction	Positive	PP	NP
	Negative	PN	NN

In the experiments, this paper adopts the accuracy rate as the criterion for judging the algorithm, for the binary classification problem, the confusion matrix is shown in Table 1, then the expression of the accuracy rate is:

$$Accuracy = \frac{PP + NN}{PP + NN + PN + NP} \tag{15}$$

Figure 1 shows the quadrant plot of 100 experimental results of regenerative kernel classification model based on Euclidean distance, as shown in Table 2, which has the highest value of accuracy of 0.68 and the lowest value of 0.48, with the mean value of 0.5895 and variance of 0.0013. Figure 2 shows the quadrant plot of 100 experimental results of the regenerative kernel classification model based on the B-spline basis function Euclidean distance, as shown in Table 2, which has the highest accuracy value of 0.68, the lowest value of 0.5, the mean value of 0.5944, and the variance of 0.00135. Figure 3 shows the quadrant plot of 100 experimental results of the regenerative kernel classification model based on functional martens distance, as shown in Table 2, which has the highest accuracy value of 0.88. And the lowest value of 0.71, with a mean value of 0.8016 and variance of 0.00096.

In summary, it can be intuitively seen from the image with the table that the algorithm proposed in this paper is superior to the other two algorithms both in terms of accuracy and degree of discretization. The reason is because the regeneration kernel constructed based on the Mahalanobis distance captures the information of the functional data, and the KPCA dimensionality reduction method is utilized to extract the nonlinear information and reduce the amount of information lost in the functional data.

Table 2: Mean and Variance of the Three Methods

	Euclidean	B-spline	functional Mahalanobis
Max accuracy	0.68	0.68	0.88
Min accuracy	0.48	0.50	0.71
Average accuracy	0.5895	0.5944	0.8016
Accuracy variance	0.001362	0.00135	0.000965

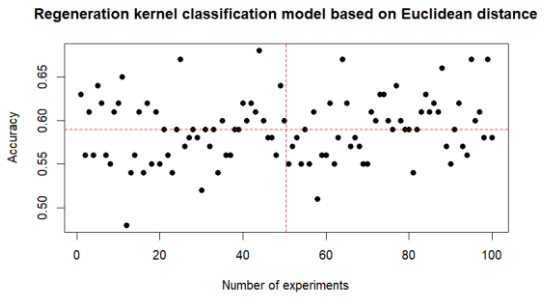

Figure 1: Results of 100 experiments based on the Euclidean distance kernel

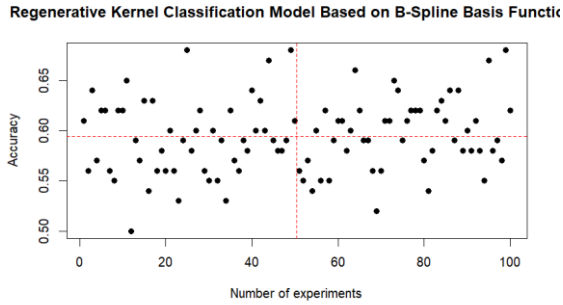

Figure 2: Results of 100 experiments based on the kernel of B-spline basis functions

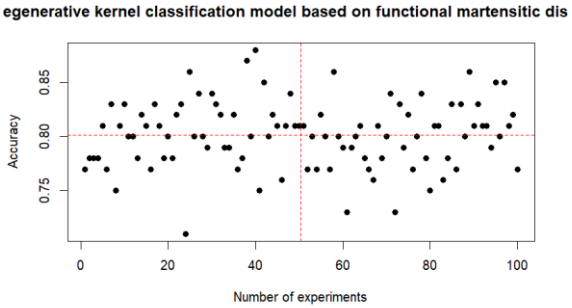

Figure 3: Results of 100 experiments based on functional Mahalanobis distance kernels

4. Conclusions and outlook

In this paper, we propose a regenerative kernel data classification model based on functional martens distance, the main idea is to convert the infinite-dimensional functional data into regular scalar data by principal component analysis through functional martens distance, and then use random forest based on KPCA to classify the data, so that it can be used to classify the functional data with any machine learning algorithm. The most important step of this algorithm is to convert the infinite-dimensional functional data into regular scalar data without losing the information of the data, and the regenerative kernel-based functional Mahalanobis distance in this paper captures the information of the functional data well, and the KPCA dimensionality reduction method can be used to extract the nonlinear information and reduce the loss of information of the functional data. Through the analysis of actual data, it shows that the regenerative kernel classification model based on function-type martens distance proposed in this paper has good competitiveness in the field of function-type data classification. For future research directions, the similarity metric proposed in this paper can be applied to other machine learning algorithms based on regenerative kernel for function-type data,

which includes solving problems such as clustering and regression for function-type data.

References

[1] Zhao Shuning. Functional support vector machines in regenerative kernel Hilbert spaces and their applications[D]. Jiangxi University of Finance and Economics,2022.

[2] Hao Sai. Research on Functional Data Classification Based on Consistency Prediction[D]. Lanzhou University, 2023.

[3] Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47(4), 379–396.

[4] Pourshoghi A, Zakeri I, Pourrezaei K. Application of functional data analysis in classification and clustering of functional near-infrared spectroscopy signal in response to noxious stimuli. J Biomed Opt. 2016 Oct;21(10):101411.

[5] Jin Haibo,Ma Haiqiang. Research on algorithms for segmented extraction of functional data features[J]. Computer Application Research,2020,37(06):1765-1768.

[6] Pini, Alessia and Simone Vantini. "Interval-wise testing for functional data." Journal of Nonparametric Statistics 29 (2017): 407 - 424.

[7] Fan Guangzhe et al. "Functional data classification for temporal gene expression data with kernel-induced random forests." 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2010): 1-5.

[8] Rossi F,Villa N.Support vector machine for functional data classification[J].Neurocomput-ing,2006,69(7):730–742.

[9] Thind, Barinder et al. "Deep Learning With Functional Inputs." Journal of Computational and Graphical Statistics 32 (2020): 171 - 180.

[10] Fuchs, Karen et al. "Nearest neighbor ensembles for functional data with interpretable feature selection." Chemometrics and Intelligent Laboratory Systems 146 (2015): 186-197.

[11] F. Rossi, B. Conan-Guez and F. Fleuret, "Functional data analysis with multi layer perceptrons," Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), Honolulu, HI, USA, 2002, pp. 2843-2848 vol.3.

[12] Wenlin D, Genton M G .Directional outlyingness for multivariate functional data[J].Computational Statistics & Data Analysis, 2018:S016794731830077X-.DOI:10.1016/j.csda.2018.03.017.

[13] Vommi Amukta Malyada and Battula Tirumala Krishna. A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study[J]. Expert Systems With Applications, 2023, 218.