A Functional Data Classification Framework Based on Functional Sufficient Dimension Reduction and Model Averaging

Qiang Wang [1, #], Pengyu Wu [2, #], Mengjie Li [3, *, #]

[1] School of Science, Harbin Institute of Technology, Weihai, China, 264299

[2] School of Mathematics and Statistics, Changshu Institute of Technology, Changshu, China, 215500

[3] School of Mathematics and Statistics, Linyi University, Linyi, China, 276000

* Corresponding Author Email: 19560957259@163.com

[#]These authors contributed equally

Abstract. With the ever-evolving advancements in data collection and storage technologies, high-frequency data recorded in patterns that change over time have become increasingly common. In many application scenarios for this type of data, functional data classification has emerged as a prominent issue in the field of statistics. In light of this, this paper proposes a functional data classification model based on the functional sufficient dimensionality reduction method and the idea of model averaging. The proposed method utilizes techniques such as Functional Slice Inverse Regression (FSIR) and Functional Average Variance Estimation (FSAVE) to project an infinite-dimensional random function onto a function space spanned by a finite wiki function, ensuring that the original data's effective information for categorical variables is not lost. Additionally, the Bagging algorithm effectively addresses the overfitting and underfitting problems that arise in single predictive models, while employing model averaging instead of model selection to adaptively select the sufficient dimensionality reduction sub-directions. Notably, in the prediction stage, the number and types of base models are flexible. Empirical analysis demonstrates that the proposed method outperforms some comparative methods in terms of prediction accuracy and robustness.

Keywords: Functional sufficient dimensionality reduction, functional data classification, Bagging algorithm.

1. Introduction

With the rapid development of global informatization, vast amounts of functional data are being generated in various fields such as medicine, finance, and industry. This study focuses on random functions as the sole data type, where the collection of sample elements with specific meanings allows for classification tasks. Such practical applications are abundant in the real world, ranging from identifying the chemical composition of substances in the field of chemistry, to understanding the pathogenic mechanisms of patients, to determining the occurrence of myocardial infarction through electrogram changes over time, to the classification of projectile points in archaeology based on the X-axis movement trajectory of handwriting symbols. As a result, this paper's research topic holds significant theoretical and practical value.

Functional data analysis involves working with high-dimensional data, which makes many traditional classification models inapplicable. There are generally two types of methods for processing functional data. The first approach involves using generalized linear models with regularization techniques. For example, Xia et al. [1] penalize the smoothness of the prediction function to deal with high-dimensional problems, while the Hu Jilei et al. [2] uses a penalty function to predict and analyze population life. The second method for processing functional data involves reducing the dimensionality of the data and then classifying it. One common approach to dimensionality reduction is principal component analysis. For instance, when Shiravani et al. [3] performed network intrusion detection, they first reduced the dimensionality of the network intrusion

feature data using principal component analysis, and then used traditional classification methods to classify the data. However, principal component analysis is an unsupervised dimensionality reduction method that is independent of predictor variables, and thus cannot guarantee the prediction accuracy of classification models. In contrast, supervised dimensionality reduction methods, such as sufficient dimensionality reduction, may have wider applicability. The Qian Jiaming et al. [4] has expanded the field of sufficient dimensionality reduction beyond the estimation of the mean dimensionality reduction space. This method enables the reduction of predictor variable dimensionality while retaining effective information. For instance, the effectiveness of this method has been verified by the study of the local influence of the objective function based on the dMAVE method in sufficient dimensionality reduction, as presented in Xu Wei [5]. Similarly, Ma Shaopei [6] demonstrates the superiority of this method by constructing a test statistic with certain dimensionality reduction properties. However, the aforementioned methods are limited to the context of multivariate data, which lacks the potential function information present in functional data. To extract this type of information, common methods involve the expansion with some basis functions, such as B-spline basis, Fourier basis, wavelet basis, and functional principal component basis. For example, Heng [7] introduces the Pearson similarity coefficient in functional data clustering analysis to address the issue of the Euclidean distance's inability to describe the morphological difference between curves. Meanwhile, However, these basis functions are fixed and unsupervised, and different types and numbers of basis functions will significantly impact the final classification outcome. By contrast, sufficient dimensionality reduction techniques for functional data appear more promising. On one hand, the problem of high-dimensional data is addressed by projecting random functions onto a function space spanned by a finite set of wiki functions, without losing valid predictive information on the corresponding variables. On the other hand, two key challenges that arise are dimension selection and model structure selection after sufficient dimension reduction. These problems can be tackled using two primary approaches, namely model selection and model averaging [8]. In contrast to model selection, model averaging can effectively measure the bias and variance of the prediction model. A classic example of a model averaging method is random forest, which has been widely used in various fields. For instance, Zhang Shiyu et al. [9] and others proposed a dimensionality reduction method based on random forest for evaluating pulse signal features. By applying the random forest algorithm and sorting the importance of each feature based on the Gini index, the algorithm selects features based on classification accuracy.

To summarize, this study proposes a functional data classification framework that employs functional sufficient dimensionality reduction techniques and the Bagging algorithm. The proposed method features two main innovations. Firstly, it uses functional sufficient dimensionality reduction to project the original data into a function space spanned by finite basis functions, enabling the feature representation in this space to retain categorical variables without loss. Secondly, the Bagging algorithm is utilized to address the issue of selecting the number of sub-directions for sufficient dimensionality reduction, and to balance the variance and bias of the prediction model. Compared to random forest, the sub-models in the weighted Bagging algorithm are unrestricted, allowing for the use of different base models depending on the complexity of the data and increasing flexibility. The results of actual data analysis confirm the high prediction accuracy of the proposed method.

2. The theory and methodology

2.1. Functional Sliced Inverse Regression and Functional Sliced Average Variance Estimates

According to literature [10], it can be seen that both available slice inverse regression and functional mean-variance estimation are generally adequate dimensionality reduction methods. For convenience, we define the following notation to denote the covariance operator and the conditional covariance operator, respectively:

$$\widehat{var}(E(x,y)) = \sum_{h=1}^{H} \bar{\chi}_h \otimes \bar{\chi}_h / H \tag{1}$$

$$\widehat{var}(E(x,y)) = \sum_{h=1}^{H} \bar{\chi}_h \otimes \bar{\chi}_h / H \tag{2}$$

Of which$\widehat{var}_h = \sum_{j=1}^{c_h} (\chi_{(h,j)} - \bar{\chi}_h) \otimes (\chi_{(h,j)} - \bar{\chi}_h)$. The steps are shown below: Firstly, on the basis of $B_1(t), \dots, B_k(t)$, the coefficients $\chi_i(t) = \chi_i^T B(t)$ can be approximated by expandingχ_i and β_j, where $\chi_i = (\chi_{i1}, \dots, \chi_{ik})^T$, $B(t) = (B_1(t), \dots, B_k(t))^T$. Suppose the matrix coefficients are $x_{n\times k} = (x_1^T, \dots, x_n^T)^T$ and $B_{k\times k} = (\langle B_i, B_j \rangle)_{i,j=1,\dots,k}$. Following on, y is divided into H slices from $I_1, \dots, I_H$. In each slice, the mean of χ_i and the covariance of the samples are calculated. This is denoted as $\overline{x_h}$ and $\widehat{cov_h}$(h = 1, … , H). Then calculate A

$$\hat{V}_{FSIR} = \alpha B^{-1}\left(\widehat{var}(x)\right)^{-1} B^{-1}\left[B\left(\widehat{var}(x)\right)B - 2B\sum_{h=1}^{H}\frac{c_h}{h}\widehat{cov_h}B + B\sum_{h=1}^{H}\frac{c_h}{h}\widehat{cov_h}\widehat{var}(x)^{-1}\widehat{cov_h}B\right] \tag{3}$$

$$\hat{V}_{FSAVE} = (1-\alpha)B^{-1}\left(\widehat{var}(x)\right)^{-1}\sum_{h=1}^{H}\frac{c_h}{h}(\overline{x_h} - \bar{x})(\overline{x_h} - \bar{x})^T B \tag{4}$$

of which c_h denotes the number of observations of the h slice, $\hat{V}_{FSIR}$ uses the first k eigenvalues of $\hat{\Gamma}^{-1}\widehat{var}(E(x,y))$ to estimate the eigenspace, and $\hat{V}_{FSAVE}$ uses the first k eigenvectors of $\Gamma^{-1}E[(\Gamma - 2var(x|y) + var(x|y)\Gamma^{-1}var(x|y)]$ to estimate the eigenspace. In addition to the above two classical function-based adequate dimensionality reduction methods, there are many other methods. For example, we can use the weighted versions of both FSIR and FSAVE methods for analysis.

2.2. The algorithm of Bagging

First of all, we will briefly describe how the Bagging algorithm works. Bagging as an integrated learning algorithm follows the following theorem for n different models $f_1(x), \dots, f_n(x)$, whose average expected error is $\bar{\mathcal{R}}(f)$. The expectation error of the integrated model $F(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x)$.F(x) based on a simple voting mechanism is between $\frac{1}{M}\bar{\mathcal{R}}(f)$ and $\bar{\mathcal{R}}(f)$. The Bagging algorithm steps are specified as follows: firstly, n subsets Di of size n' are selected from them uniformly and with put-back (i.e., using the self-sampling method) as new training sets. Then, the base model is used on these n training sets using classification and regression algorithms. Then n models can be obtained. Finally, the results of Bagging can be obtained by classifying them through the voting method. The specific flow chart is shown in Figure 1:

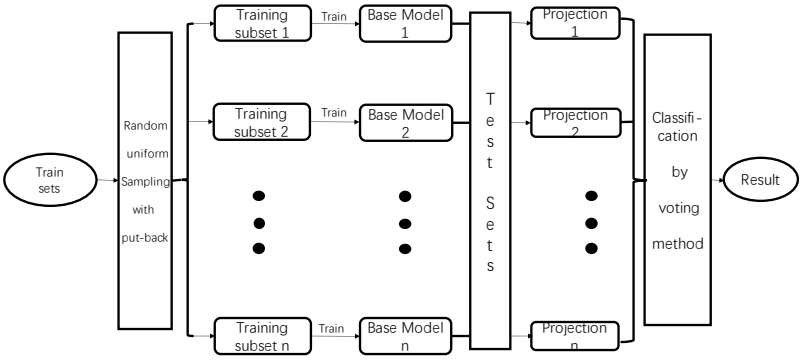


Figure 1. The algorithm of bagging

Next, we will propose a functional data classification model based on a functional sufficient downscaling and Bagging algorithm with the following steps:

Step1: Divide the data set into a training set, a validation set, and a test set, where the training set is used to generate the model parameters, the validation set is used to adjust the hyperparameters, and the test set is used to test the effectiveness of the algorithm.

Step2: Generate projection function vectors on the training set according to the category bins and using the weighted versions of FSIR and FSAVE and apply them to the projection coefficient matrices

of the validation and test sets to generate the corresponding projected training, validation, and test sets.

Step3: Bring the data set into the Bagging algorithm, get the prediction results of each sub-model and vote on them, and output the final prediction results.

Step4: The proposed method is compared with the comparison method in 100 simulation experiments, and the mean and standard deviation of 100 simulation experiments are specifically calculated to test the effectiveness of the proposed method.

3. Data analysis

3.1. Data source and experimental description

The dataset used in this article was obtained from the public data source UCR (University of California, Riverside), consisting of Three datasets: TwoLeadECG, GunPointMaleVersusFemale, and Strawberry. Specifically, TwoLeadECG is a dataset collected and organized by UCR for electrocardiogram (ECG) signals containing 1,162 ECG records related to two-lead ECG. This dataset is commonly used to compare the performance of different algorithms on ECG signal classification, aiming to assist medical researchers and engineers to explore how to use machine learning algorithms for automatic identification of arrhythmias in ECGs, to better assist clinical doctors in making accurate diagnosis and treatment decisions. UCR's GunPointMaleVersusFemale dataset consists of 406 samples, formed by accelerometer data on hand movements during gun actions performed by individuals of different genders. It is widely used to study gender differences' impact on human action recognition by comparing male and female models' gun training sequences. It has been extensively applied in fields such as human motion recognition, machine learning algorithm assessment, and model comparison. UCR's Strawberry dataset is a benchmark database of time-series classification problems. The original data of the dataset were collected from a strawberry plantation in southern Spain, covering 983 sample data from November 2016 to June 2017. Each time series represents quality indicators of strawberries, such as weight, diameter, color, hardness, etc., within an hour. This dataset is widely used to evaluate time-series classification algorithms and has become one of the standard benchmarks in the research field of time-series data analysis.

The aim of this experiment is to use a function-based data classification model based on functional dimensionality reduction and Bagging algorithms to classify the data features of the Three datasets: TwoLeadECG, GunPointMaleVersusFemale, and Strawberry. The performance evaluation indicators of this model will be compared and analyzed with traditional machine learning algorithms such as SVM, RF, KNN, based on PCA dimensionality reduction to determine the superiority of the proposed model in this paper. The experimental design follows the following methodology:

Data pre-processing: To ensure that the input data of the model is comparable, eliminate the dimensional differences between different features, and make the data easier to compare and analyze, we standardized all data samples. We used the sklearn library in Python to randomly split each dataset into 70% for training and 30% for testing.

Feature extraction: We performed weighted FSIR-FSAVE dimensionality reduction based on FPCA and PCA dimensionality reduction on the divided datasets, respectively, to obtain two sets of low-dimensional feature data that can effectively reflect the original data information.

Model design: Bagging algorithm was used for data that was dimensionally reduced using weighted FSIR-FSAVE based on FPCA, and SVM was chosen as the base model with 100 base models. For data that was dimensionally reduced using PCA, traditional machine learning algorithms were used. The F1 value and accuracy ACC were calculated for the classification results obtained by the two different dimensionality reduction and classification algorithms. The formula for F1 value is as follows:

$$F_1 = \frac{2TP}{2TP+FP+FN} \tag{5}$$

The formula of accuracy ACC is as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{6}$$

The terms TP,TN,FPand FN represent the quantities of correctly classified positive instances, correctly classified negative instances, misclassified positive instances, and misclassified negative instances, respectively. To avoid the effects of random chance in the data set partitioning process, we repeated this procedure 100 times and generated an F1 value and accuracy (ACC) box plot, along with key model evaluation metrics such as the mean, variance, and bias. Finally, we visually compared and analyzed the efficacy of the Bagging algorithm and the traditional machine learning algorithms using FPCA-based weighted FSIR-FSAVE, ultimately reaching a conclusion based on our model evaluation metrics.

3.2. Analysis ofcomparative results

By using Python, we constructed models and performed adaptive tuning. After running 100 times on randomly split data sets, we obtained the corresponding F1 scores and ACC accuracies for different models, as shown in Table.1, Table.2, and Table.3:

Table.1. The relevant metrics obtained from the TwoLeadECG dataset.

Model	PCA-lr	PCA-knn	PCA-tree	PCA-svm	PCA-RF	Bagging-FSDR
F1 mean value	0.5999	0.7412	0.6878	0.7845	0.7410	0.9672
F1 variance value	0.0005	0.0003	0.0006	0.0003	0.0003	0.0001
F1 Standard variance value	0.0224	0.0196	0.0245	0.0185	0.0187	0.0096
ACC mean value	0.5915	0.7318	0.6914	0.7575	0.7238	0.9674
ACC variance value	0.0004	0.0004	0.0005	0.0004	0.0003	0.0001
ACC Standard variance	0.0209	0.0190	0.0215	0.0195	0.0184	0.0093

Table.2. The relevant metrics obtained from the GunPointMaleVersusFemale dataset.

Model	PCA-lr	PCA-knn	PCA-tree	PCA-svm	PCA-RF	Bagging-FSDR
F1 mean value	0.5571	0.8943	0.8758	0.8643	0.8770	0.9703
F1 variance value	0.0037	0.0010	0.0011	0.0010	0.0011	0.0002
F1 Standard variance value	0.0613	0.0321	0.0335	0.0325	0.0342	0.0168
ACC mean value	0.5631	0.8896	0.8698	0.8537	0.8697	0.9680
ACC variance value	0.0025	0.0010	0.0011	0.0009	0.0011	0.0003
ACC Standard variance	0.0502	0.0317	0.0335	0.0305	0.0336	0.0174

Table.3. The relevant metrics obtained from the Strawberry dataset.

Model	PCA-lr	PCA-knn	PCA-tree	PCA-svm	PCA-RF	Bagging-FSDR
F1 mean value	0.4112	0.8330	0.8005	0.8119	0.8438	0.8616
F1 variance value	0.0019	0.0006	0.0010	0.0007	0.0007	0.0005
F1 Standard variance value	0.0436	0.0237	0.0313	0.0269	0.0260	0.0213
ACC mean value	0.6537	0.8769	0.8564	0.8601	0.8841	0.8974
ACC variance value	0.0006	0.0003	0.0005	0.0004	0.0003	0.0002
ACC Standard variance	0.0252	0.0175	0.0212	0.0194	0.0186	0.0157

Tables.1 through Tables.3 demonstrate that the weighted FSIR-FSAVE algorithm based on FPCA and Bagging outperforms other traditional machine learning classification algorithms in terms of the mean, variance, and standard deviation of F1 scores and ACC accuracies. This indicates that the model constructed in this paper has extremely high accuracy, stability, and reliability on these three data sets. Additionally, compared with other models, the weighted FSIR-FSAVE algorithm based on FPCA and Bagging exhibits superior data fitting and prediction capabilities. Furthermore, this paper uses kernel density plots to visually represent the probability density of data values within a continuous numerical range, as well as the distribution of F1 scores and ACC accuracies, including information such as peak position, quantity, and height. Figures 2 to 4 display the kernel density plots of F1 scores for different models under different data sets, while Figures 5 to 7 display the kernel density plots of ACC accuracies for different models under different data sets.

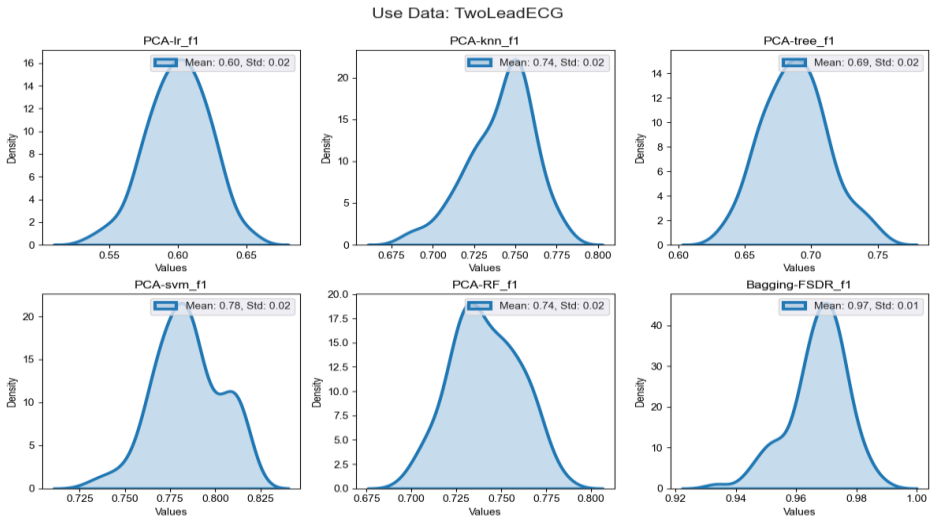
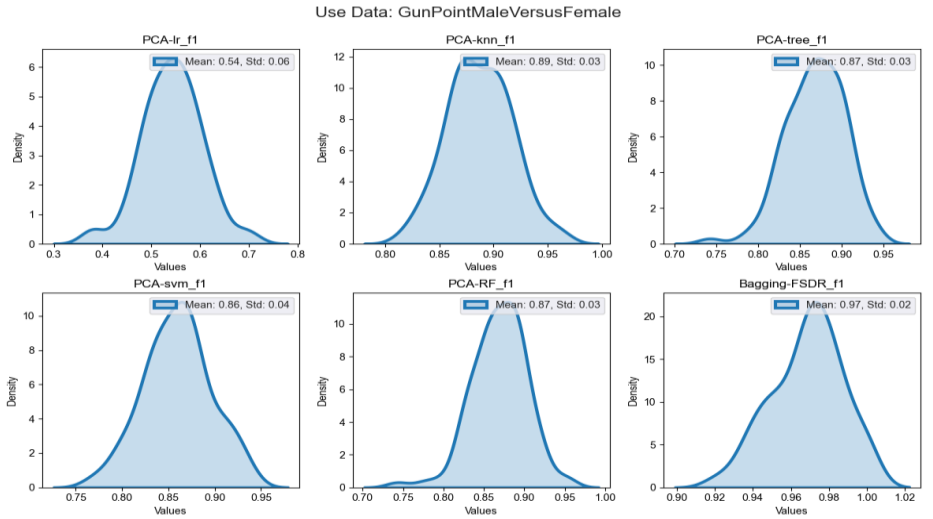
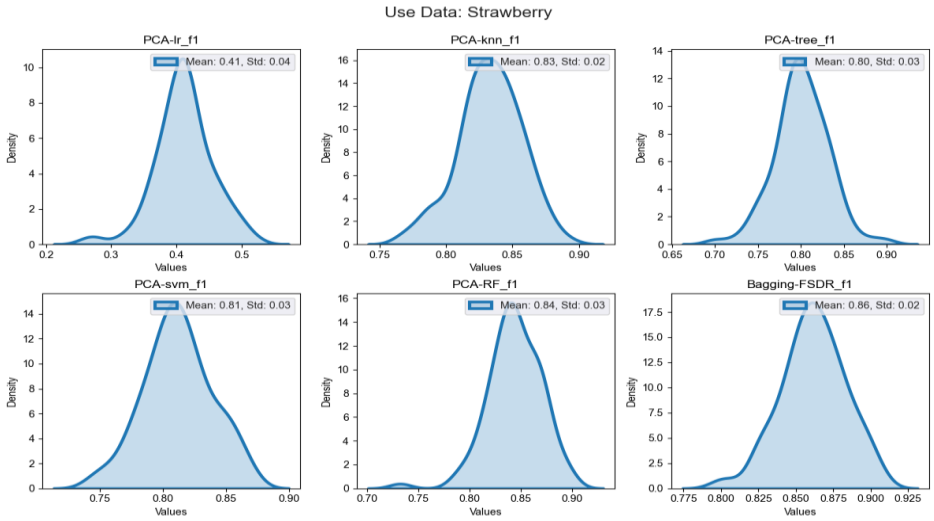


Figure 2. F1-Nuclear density map of 100 experiments(TwoLeadECG)



Figures 3. F1-Nuclear density map of 100 experiments (GunPointMaleVersusFemale)



Figures 4. F1-Nuclear density map of 100 experiments (Strawberry)

Based on the kernel density estimation plots of F1 scores for different models under three datasets, we can observe that the peak of the weighted FSIR-FSAVE algorithm based on FPCA and Bagging is higher than that of other traditional machine learning algorithms, with a sharper peak. This indicates that the proposed model has a concentrated distribution and the highest frequency of F1 scores among the three datasets. Additionally, the kernel density curves of this model exhibit a left-skewed distribution, suggesting a positive skewness of F1 score data. Therefore, we conclude that our model shows good stability and generalization ability.

In Figures 5 to 7, we observe that the distribution of accuracy (ACC) scores for the weighted FSIR-FSAVE algorithm based on FPCA and Bagging is more concentrated than that of other traditional machine learning algorithms. Furthermore, this model exhibits better data fitting and prediction capabilities. Additionally, we employed box plots to illustrate the distribution of F1 scores and ACC accuracy for corresponding 100 experiments, including median, quartiles, minimum and maximum values, as well as outliers, as shown in Figures 8 to 10.

From these six box plots, we can observe that the F1 scores and ACC accuracy obtained by the weighted FSIR-FSAVE algorithm based on FPCA and Bagging are higher than those of traditional machine learning algorithms under all three datasets, particularly for the TwoLeadECG and GunPointMaleVersusFemale datasets. Furthermore, compared to traditional machine learning algorithms, the F1 scores and ACC accuracy of our proposed model are more concentrated, and the number of outliers is relatively small. The corresponding maximum, median, quartiles, and minimum values are also higher than those of traditional machine learning algorithms. This further demonstrates that the weighted FSIR-FSAVE algorithm based on FPCA and Bagging exhibits a balanced performance for classifying samples of different categories and outperforms traditional machine learning algorithms to some extent.

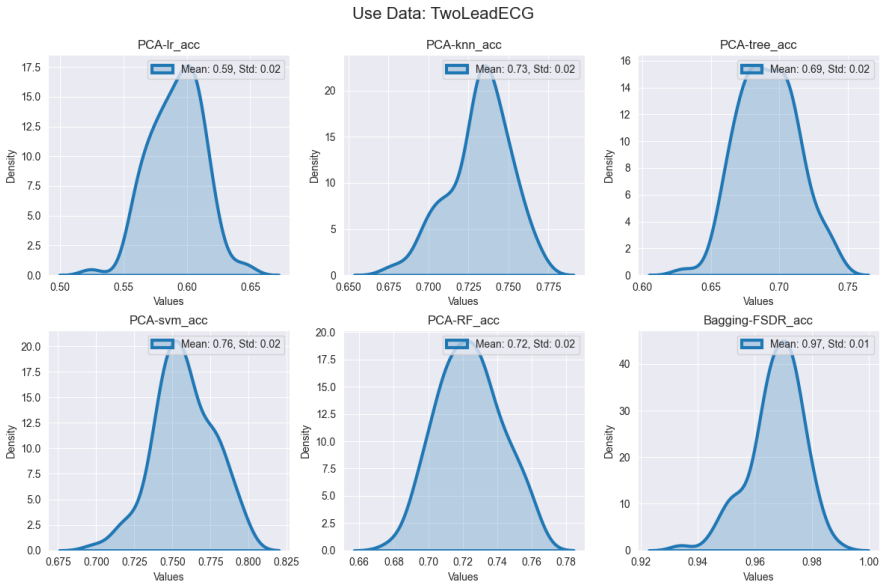


Figure 5. ACC-Nuclear density map of 100 experiments (TwoLeadECG)

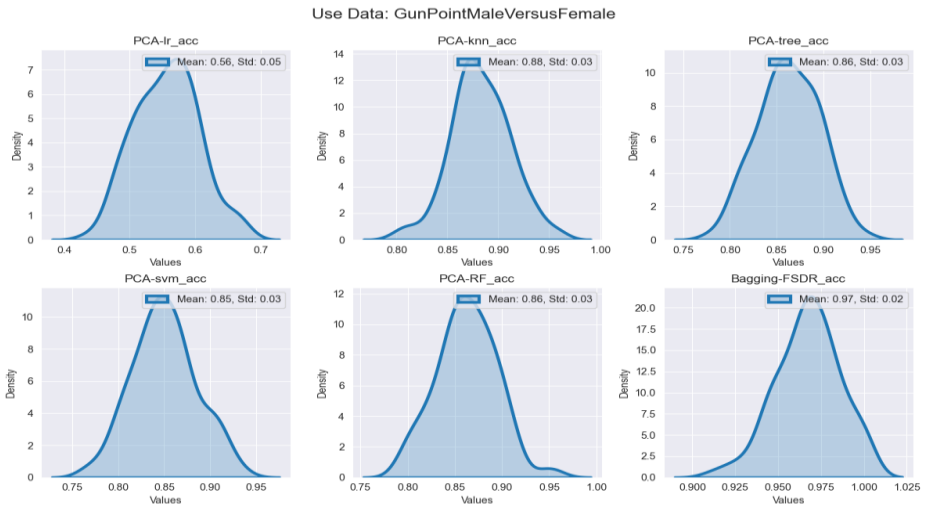


Figure 6. ACC-Nuclear density map of 100 experiments (GunPointMaleVersusFemale)

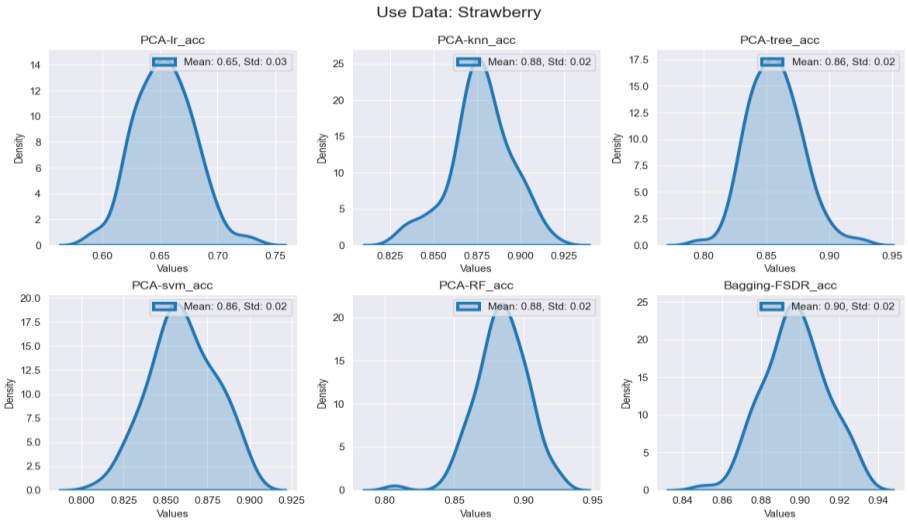


Figure 7. ACC-Nuclear density map of 100 experiments (Strawberry)

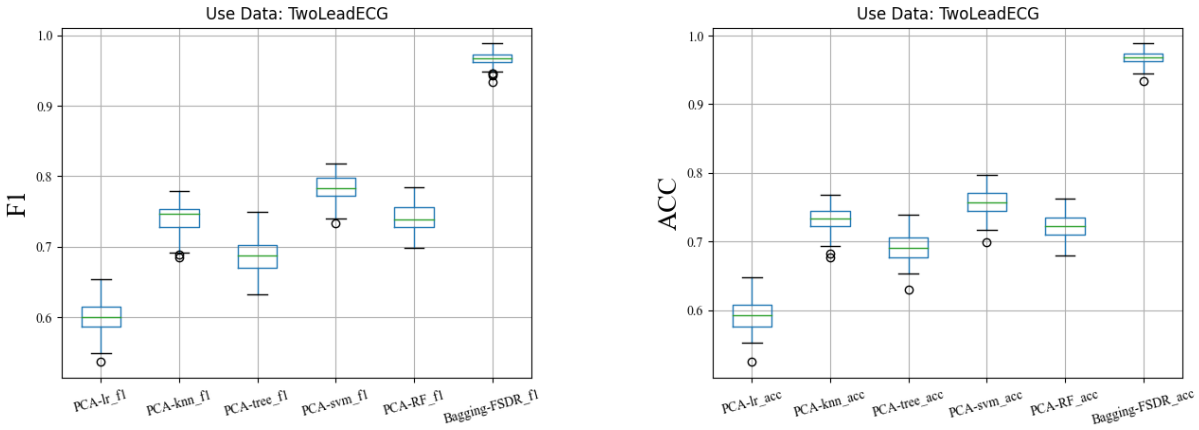


Figure 8. F1 and ACC boxplot of 100 experimental results(TwoLeadECG)

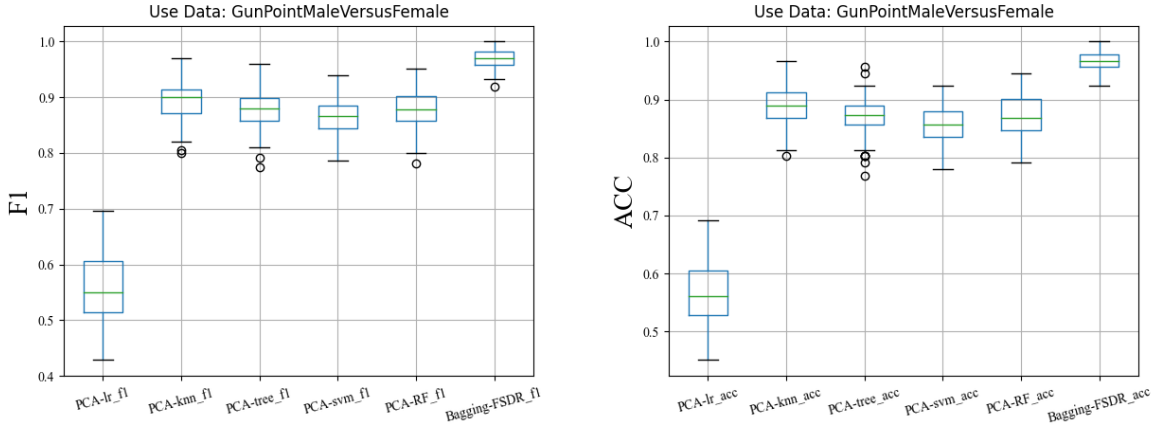


Figure 9. F1 and ACC boxplot of 100 experimental results(GunPointMaleVersusFemale)

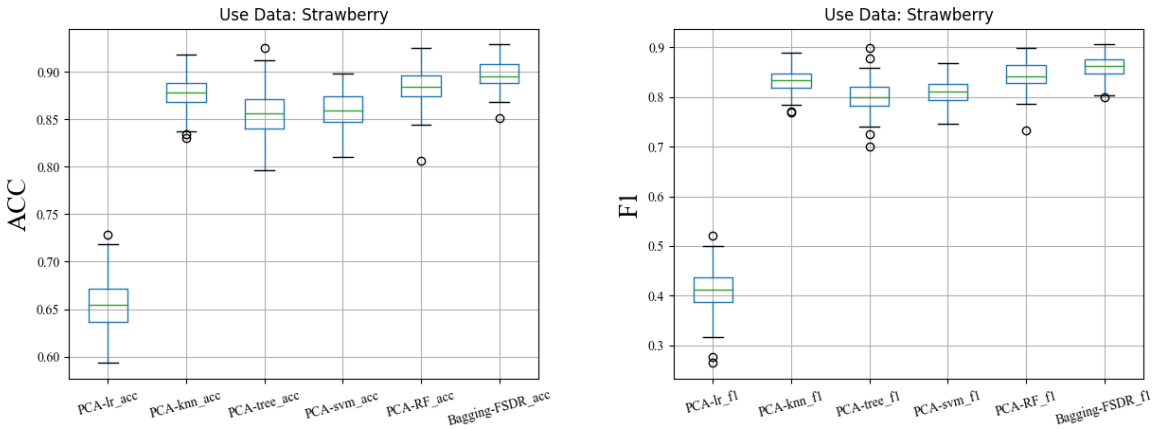


Figure 10. F1 and ACC boxplot of 100 experimental results(Strawberry)

4. Conclusions

This paper mainly proposes to construct a functional data classification framework based on functional sufficient dimensionality reduction and model averaging. The core direction is to project infinite-dimensional random functions to finite Wiki function space to improve prediction accuracy and robustness. The construction of the model framework is conducive to solving the variance and bias of the trade-off prediction model when dealing with a large number of functional data in real problems. When building the model, we chose adaptive parameter tuning, randomly divided the data sets in different models and ran them 100 times to obtain the F1 value and ACC accuracy rate. The

experimental data images show that the F1 value and ACC value accuracy of the model constructed in this paper are relatively concentrated. Strong, the maximum value, median, quartile, and minimum value are superior to traditional machine algorithms when the number of outliers is relatively small. Among them, the algorithm based on FPCA weighted FSIR-FSAVE and Bagging can handle different types of sample classification in a balanced manner and has certain advantages over traditional machine learning algorithms. For future work, the current functional sufficient dimensionality reduction is more meaningful to solve the sample problem on some high-dimensional data. On the surface, the function model mainly deals with dense functional data, but for sparse functional data with relatively low dimensionality, We further optimize the current functional data classification model so that it can still maintain high accuracy and good fitting effect when dealing with sparse functional data.

References

[1] Xia Mengyao,Cai Helen Huifen. The driving factors of corporate carbon emissions: an application of the LASSO model with survey data .[J]. Environmental science and pollution research international,2023.

[2] Hu Jilei, Wu Wenliang, Wang Jing et al. Logistic regression discriminant model for sand liquefaction based on adaptive LASSO [J]. Journal of China Three Gorges University (Natural Science Edition),2023,45(02):67-72.

[3] Shiravani Anita,Sadreddini Mohammad Hadi,Nahook Hassan Nosrati. Network intrusion detection using data dimensions reduction techniques [J]. Journal of Big Data,2023,10(1).

[4] Qian Jiaming,Cao Yu,Bi Ying,Wu Hongjun,Liu Yongtao,Chen Qian,Zuo Chao. Structured illumination microscopy based on principal component analysis [J]. eLight,2023,3(1).

[5] Xu Wei. Local impact analysis based on objective function under dMAVE method in sufficient dimensionality reduction [D]., Yunnan University of Finance and Economics 2022.

[6] Ma Shaopei, Tian Maozai. Heteroscedasticity Test for Semiparametric Multi-Indicator Models Based on Partial Sufficient Dimensionality Reduction Method [J]. Chinese Science: Mathematics,2022,52(08):935-968.

[7] Heng Lian. Functional sufficient dimension reduction: Convergence rates and multiple functional case [J]. Journal of Statistical Planning and Inference,2015,167.

[8] Liao Jun, Wen Li, Yin Jianxin. Selection and average estimation of high-order spatial autoregressive models [J]. Systems Science and Mathematics,2021,41(05):1400-1417.

[9] Zhang Shiyu, Yang Ke, Xia Chunming, etc. Dimensionality reduction and classification of pulse signal features based on random forest [J]. World Science and Technology - Modernization of Traditional Chinese Medicine,2020,22(07):2418-2426.

[10] Guochang Wang, Yan Zhou, Xiang-Nan Feng, Baoxue Zhang. The hybrid method of FSIR and FSAVE for functional effective dimension reduction.