A Novel Divide and Merge Approach for Improved Classification of Functional Data

Wei Zhao, Xiao-jun Zeng*, Chengdong Shi, Ching-Hsun Tseng, Yue Chang
University of Manchester

Introduction

Functional data classification is one of the key technologies for analyzing complex data structures with inherent continuity and plays a crucial role in various fields. Although traditional classification methods such as logistic regression, decision tree, and support vector machines (SVM) have achieved some success in processing such data, they often fail to fully capture the continuity and inherent characteristics of the data. In recent years, with the increase in functional data applications [1], researchers have begun to explore more refined data approximation methods based on FDA [2] to improve the accuracy and efficiency of functional data classification. By using FDA, the raw data points are transformed into curves rather than discrete vectors, emphasizing the function nature of the data, and it allows a more comprehensive understanding and representation of the data. As FDA considers data in their continuous form, it shows the underlying patterns and trends within the data, enhancing the analysis and subsequent classification processes.

The main contribution of this work are as follows:

- We proposed the divide-merge method to find a right set of knots and then construct a set of common basis functions to represent all the functions in a raw dataset.
- Through the common basis functions we got, the traditional high-dimensional learning problems are converted from an infinite-dimensional vector space to a corresponding finite-dimensional parameter space, which can be solved by traditional machine learning methods, thus reducing dimensionality, enhancing simplicity and improving efficiency.
- Through several experiments, it is shown that the approach let to simpler classifiers and improved classification accuracy, thereby enhancing the model's interpretability and performance, confirming the feasibility and effectiveness of the method.

Methods

Divide:

In this step, we divide the data according to labels. We assume that all data from same label share some common features, therefore, the mean series composed of the mean values of every observation point with the same label were treated as that label's representative. From a function point of view, it shows the overall shape of one kind of input data.

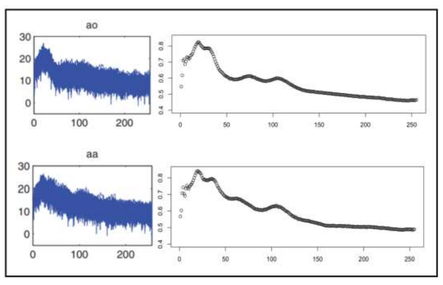

Fig 1. The mean series of the two labels.

Knot placement:

To serve to our purpose, the Fast Automatic Knot Placement method [3] is applied and extended to be used in this step.

We calculate the knot placement of the two mean series in Fig.1. The knot vector from 0 to 80 which is the most different part between data 'aa' and 'ao' is displayed in Fig.2. The triangle shows the knot in the mean series.

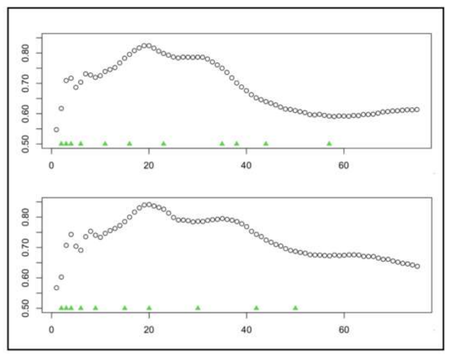

Fig 2. Knot placement for two mean series.

Knot merge:

Let K_i be the knot of vector for mean series of category i, $d_{ij} = K_i[j+1] - K_i[j]$ represents the distance between adjacent knots in category i, and $\overline{d}_i$ be the average distance between adjacent knots for category i:

$$\overline{d}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_{i-1}} d_{ij} \quad (1)$$

Then, use half of the average of these averages as a threshold.

$$Threshold = \frac{1}{2m} \sum_{i=1}^{m} \overline{d}_i \quad (2)$$

And the final common knot vector should be:

$$K_{final} = sort(\{K_i[j] | d_{i,j} < Threshold\}) \quad (3)$$

Smoothing:

Given the knot vectors merged from above, when the common B-spline basis functions are used, the only variable $\{P_i\}$ decide the shape of the B-spline curve. Then the coefficients $\{P_{it}\}$ of B-spline basis function, is set to represent the t-th input function in parameter space.

A set of n common B-spline basis functions $B_i(x)$ were used to fit the raw data. Assume T input curves, for each discrete observation $(x_{jt}, y_{jt})(j = 1..m, t = 1..T)$, where x_{jt} represents the independent variable, and y_{jt} the dependent variable. y_{jt} can be expressed as follows:

$$y_{jt} = \sum_{i=1}^{n} P_{it} B_i(x_{jt}) + \varepsilon_{jt} \quad (4)$$

The least squares method is utilized to calculate P_{it}. With the coefficient vector {Pik}, the curve fitted by B-spline can be calculated, and one random "ao" fitting result is shown in Fig.3.

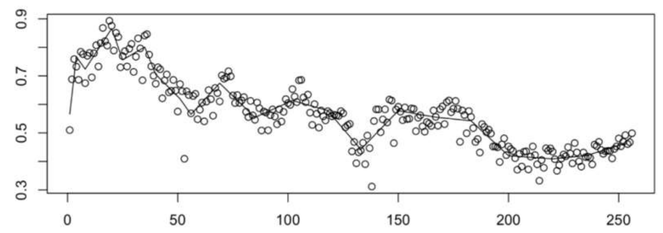

Fig 3. Raw data point and the corresponding continues curve.

Classification:

In general, the challenging task of defining classifier F in the infinite-dimensional input spaces is simplified by converting it into a manageable learning problem for the model f , which is a model in the finite-dimensional spaces of input parameters. Since the model f 's learning problem exists in a traditional vector space, the classification issue can be addressed using the existing machine learning method, such as SVM or k-NN.

Results

Phoneme Dataset:

The smoothing of one random normal sample "ao" is shown in Fig. 3. And performance in Table.1.

	Divide&Merge +SVM	RKVS +QDA[4]	RF[4]	k-NN[4]
Phoneme	**0.930±0.004**	0.927±0.005	0.924±0.006	0.911±0.004
Duo	**0.819±0.030**	0.815±0.015	0.810±0.015	0.803±0.015
Tecator	0.989±0.019	0.978±0.015	**0.990±0.012**	0.980±0.019
Fish	**0.894±0.056**	0.824±0.037	0.806±0.037	0.776±0.033

Table 1. Classification results.

Tecator Dataset:

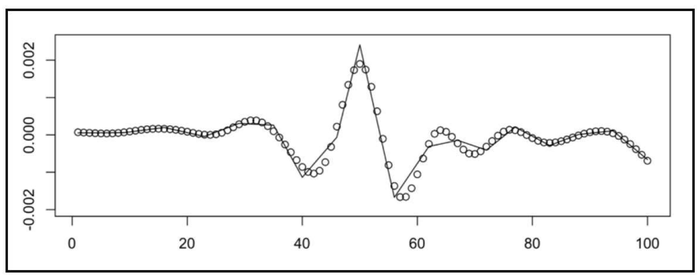

Fig 4. The observation and fitted curve for one sample with high fat content.

Fish Dataset:

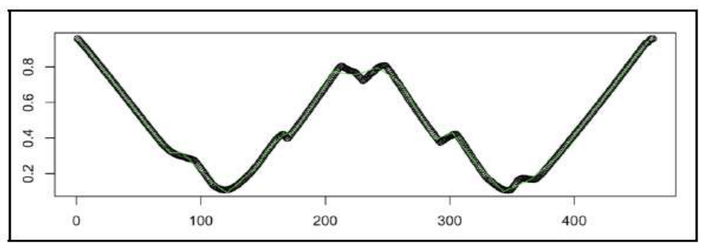

Fig 5. The smoothed result for one random picked sample.

Medical Image Dataset:

Divide & Merge + SVM (10 nbasis)	Diststace[5]
0.728±0.035	**Around 0.70**

Table 2. Classification result for MRI image.

Conclusions

In this paper, we proposed a novel feature representationWe proposed a divide and merge method for functional data classification based on FDA. At the beginning, we divided the dataset according to the label and identify the knot vector for each dataset with the same label, and then apply the knot merge to build the knot vector for whole dataset with all labels in classification problems. Then the merged knot vector accomplished the fitting of B-spline for the raw data, and least square method was utilized to calculate the coefficient vector P in B-spline curve fitting. This function representation step encodes the raw data into a parameter space, which is smaller than the original input space. Subsequently, conventional machine learning techniques were used to address the classification problems within this parameter space. We tested our method on 4 real data sets. The fitted curves were displayed and illustrate we use the coefficient vector P calculated by merged knot vector to capture the function nature of these datasets. And the accuracy classification results show advantages in multiclass datasets (Phoneme, Fish, Medical Image).

*Corresponding Author: Xiao-jun Zeng

Contact

Wei Zhao
University of Manchester
Email: wei.zhao-6@postgrad.Manchester.ac.uk

References

1. Shahid Ullah and Caroline F Finch. Applications of functional data analysis: A systematic review. BMC medical research methodology, 13:1–12, 2013
2. J. Ramsay and B.W. Silverman. Functional Data Analysis. Springer Series in Statistics. Springer New York, 2006. ISBN 9780387227511.
3. Raine Yeh, Youssef SG Nashed, Tom Peterka, and Xavier Tricoche. Fast automatic knot placement method for accurate b-spline curve fitting. Computer-aided design, 128:102905, 2020
4. Carlos Ramos-Carreno, Jose Luis Torrecilla, and Alberto Suarez. Classification of functional data: A comparative study. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pages 866–871. IEEE, 2022.
5. Mia Hubert, Peter Rousseeuw, and Pieter Segaert. Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification, 11: 445–466, 2017.