Dynamic Functional Bandwidth Kernel-based SVM: An Efficient Approach for Functional Data Analysis

Anima Pramanik, Vikram Nande, Arka Shankar Pradhan, Sobhan Sarkar*, and J. Maiti

Abstract Functional data analysis (FDA) refers to the study of data having more number of features than the observations and a function of time. As functional data is based on time, one need to get the optimized values of functional bandwidth kernel over time, thereby forming continuous optimization problem. Support vector machine, abbreviated as SVM, with functional bandwidth kernel is used as a benchmark tool to classify the functional data. In the state-of-the-art analyses, the determination of the number of functional bandwidth kernels is user-defined; thereby making the process static. In this study, the aforesaid issue is solved by defining the number of functional bandwidth is as same as the number of attributes, thereby making the process dynamic in nature. In addition, functional bandwidth parameters are updated using the backpropagation of hill-climbing loss to increase the classification accuracy. An extensive study is done over various benchmark datasets to exhibit the efficacy of our proposed method over the other state-of-the-art methods.

Anima Pramanik
Department of Industrial & Systems Engineering, IIT Kharagpur, India.
e-mail: apramanik17@gmail.com

Vikram Nande
Department of Mechanical Engineering, IIT Kharagpur, India.
e-mail: vikramnande@gmail.com

Arka Shankar Pradhan
Department of Industrial & Systems Engineering, IIT Kharagpur, India.
e-mail: arkashankar2003@gmail.com

Sobhan Sarkar
Division of Management Science, Business School, The University of Edinburgh, UK.
e-mail: sobhan.sarkar@gmail.com; sobhan.sarkar@ed.ac.uk
**Corresponding author*

J. Maiti
Department of Industrial & Systems Engineering, IIT Kharagpur, India.
e-mail: jhareswar.maiti@gmail.com

Key words: Functional data analysis; Continuous optimization problem; Functional bandwidth kernel; Hill climbing loss; Optimization.

1 Introduction

Functional data analysis (FDA) has gained a large attention in various fields, such as speech recognition [1], meteorology [2], client segmentation [3], etc. FDA refers to the standard multivariate analysis over the data having infinite dimension. However, problem occurs in analysis when the number of observation is less than the number of features present in data. In addition, the features have correlation, which is not considered in multivariate analysis that leads to difficulty in handling the functional behavior of data. In order to solve this issue, one of such recent research work by [4] provides a solution for FDA in functional behavior classification, i.e., the way by which functional data (FD) is to be classified into two given classes with the use of information of training samples. Support vector machine (SVM), a popular classification algorithm in machine learning paradigm [5, 6, 7, 8, 9, 10], is considered to solve this problem due to its wide application over FD and multivariate data analysis [11]. In this study, a new functional kernel is developed, which weighs optimally the function values. The number of functional bandwidth used in this study is user-defined. Moreover, a hinge loss function has been utilized with SVM, which can be replaced by more standard loss function, i.e., ramp loss.

For SVM, [1] utilizes the functional behavior of data by converting classical kernel to functional kernel through transformation-based kernel. SVM algorithm has a kernel which is a mathematical function. The kernel takes input as data and outputs the required form. There are many kinds of kernel functions. RBF kernel function is the mostly used of them due to its simplicity [4, 12, 13, 14, 5, 15]. Therefore, we have also considered RBF kernel in this study. The conventional kernel of SVM is converted to functional bandwidth kernel, and used in FD analysis. Functional bandwidth kernel is the function of time variant data. Therefore, one need to optimize the value of functional bandwidth kernel over time, which is called continuous optimization problem [4]. Instead of multi classification error rate, we use correlation score between the class and SVM score to solve the continuous optimization problem [16].

Based on the above-mentioned discussion, though in brief, the following issues can be identified:

(i) The number of functional bandwidth used in this study is user-defined.
(ii) Loss function has not been modified to increase the classification performance

Based on these research issues, the present study contributes in the following ways:

(i) The number of functional bandwidth parameters are considered same as that of the number of attributes. In each iteration of the optimization process. Therefore, the number of values for all bandwidth parameters remains same before and after the optimization, which makes the overall process dynamic.
(ii) Functional bandwidth parameters are updated using the back propagation of hill climbing loss to increase the classification accuracy. The approach is that

it moves to the best solutions or move towards the top the hill which is local search algorithm. For all bandwidth parameters, this method is used to get optimal value of those parameters for each attribute.

The paper is structured as follows: In Section 2, the proposed methodology is given. In Section 3, the experimental results are reported and discussed, and finally, the conclusion with limitations and scope for future works are presented in Section 4.

2 Methodology

In this study, SVM is modeled for classification of FD. Sample observation is considered to be s and each observation i belongs to s, and s has a pair of (X_i,Y_i), X_i is the i-th row of the dataset. Furthermore, $Y_i \varepsilon \{-1\ ,\ +1\}$ indicates the label of class for i-th observation. If a kernel $K : X \times X \Rightarrow R$ be considered in SVM, then, we can calculate the score $\hat{Y}$ in of Eq. (1):

$$\hat{Y}(X) = \sum_{i \varepsilon s} \alpha_i Y_i K(X, X_i) \tag{1}$$

Here, α_i is obtained as a optimal solution by solving given optimization problem, as shown in Eq. (2):

$$\begin{aligned} max_\alpha \sum_{i \in s} \alpha_i - \frac{1}{2} \sum_{i,j \in s} \alpha_i \alpha_j Y_i Y_j K(X_i, X_j) \\ s.t. \sum_{i \in s} \alpha_i Y_i = 0 \\ \alpha_i \in [0, C], i \in s \end{aligned} \tag{2}$$

The classification follows in the way that if the observation has $\hat{Y} > \beta$, then, it is allotted $+1$; otherwise, -1. Here, β denotes a threshold value. A grid search with k-fold cross-validation is used for tuning scalar regularization parameter on sufficiently large intervals. The Gaussian kernel used in this study for FD can be expressed by Eq. (3) [17, 18]:

$$K(X_i, X_j) = exp(-\sum_{t=1}^{d} (X_{it} - X_{jt})^2 \omega), X_{it}, X_{jt} \in R^d \tag{3}$$

In those studies, scalar values are used as the bandwidth for the kernel. In our experiment, we have used functional bandwidth $\omega(t)$ instead of scalar fixed bandwidth ω, as expressed in Eq. (4):

$$K(X_i, X_j) = exp(-\sum_{t=1}^{d} ((X_{it} - X_{jt})^2 \omega_t)), X_{it}, X_{jt} \in R^d \tag{4}$$

We considered each bandwidth which transforms itself to the shape and structure of data which results in improved performance. In particular, obtaining optimal ω relies on t. It also provides information of those sub-intervals in $[0, T]$ which re-

mains crucial in classification, particularly, in such situation, where $\omega(t)$ becomes maximum. We have vector θ used for parameterization of certain classes of function and the vector θ belongs to certain set ϕ. Now, ω is expressed as $\omega(\theta)$. Therefore, $\omega = \{\omega_1, \omega_2, ..., \omega_d\}$, and $\theta = \{\theta_1, \theta_2, ...\theta_H\}$, where H implies a user-defined number of initial solution. The proposed method is explained step-wise as follows:

Step 1- Grid search is applied to find parameter C which is the scalar regularization parameter. Once we obtain the pair of values (C, θ), one first solves Eq. (2) so that we can get the value of the coefficient α of the score function as Eq. (1). We obtain the optimal values of θ by maximizing the correlation (here, Pearson correlation) between the class (Y_i, X_i), and the score $\hat{Y}(X_i, \theta, \alpha)$.

Step 2- If we evaluate our classifier over same data-set, this will lead to over-fitting. We avoid over-fitting by splitting the data set using k-fold cross-validation. We divide the data-set into training, testing, and validation sets. In our case, we create four independent samples, namely s_1, s_2, s_3, and s_4. The samples, s_1, and s_2 are used for training, whereas, the samples, s_3, and s_4 are used for testing, and validation, respectively.

Step 3- The sample s_1 is used to solve the Eq. (2) to obtain α. We used the sample s_2 for measuring the quality of parameter θ and C. It is used to calculate correlation between score and the class labels.

Step 4- The sample s_3 is used for calculating the accuracy and to find the regularization parameter C for all possible values in grid. We select C which provides the maximum accuracy. The final work is to calculate the accuracy in the independent sample s_4.

Step 5- The bi-level optimization problem for solving θ can be formulated as Eq. (5):

$$\begin{aligned} & max_{\theta} R(Y_i, YhX_i, \theta, \alpha))_{i \varepsilon s_2} \\ & s.t.\ \ \alpha \text{ solves } Eq.(2) \text{ in } s_1 \\ & \qquad \theta \varepsilon \Phi \end{aligned} \tag{5}$$

The above equation is a non-linear optimization problem. This can be solved by the process adopted in [19]. The summary is that the value of C is chosen from grid and then for every value of C we measure the accuracy in s_3. The C with the best accuracy is chosen. Finally, s_4 is used for estimation of classification rule. The pseudo-code of the heuristic is shown below.

3 Results and Discussions

3.1 Experimental setup: In this study, the proposed algorithm is coded in Python (using Spyder 3.2.6) and run using a 2 TB RAM. The system used here is 64-bit Windows 10 operating system with 64-bit processor. Table 1 shows the maximum accuracy obtained for $h = 2, ..., 5$, which exceeds the accuracies obtained by k-NN and classic/conventional SVM method. The reason for this improvement may be related to the shape possessed by the curves. Different types of class labels are easy to be identified by the proposed method and it seems to easily identify the sub-intervals, as well. In 'breast cancer' dataset, the improvement in accuracy is consid-

Algorithm 1: Heuristic for parameter tuning.

Input:-H,φ Where $\varphi = \{\theta_1, \theta_2, ..., \theta_d\}$and $\theta_i = \{\omega_{i1}, \omega_{i2}....., \omega_{id}\}$
Randomly split the sample s into s_1 , s_2 , $s_3$and s_4
for all c in C **do**
  Randomly select l ε [0,h]
  $\theta = \phi$
  Solve for α_{opt} using problem (2) on s_1
  Solve for θ_{opt} using problem (5) on s_2
  Randomly select l ε [0,h]
  for i in $len(s_3)$ **do**
    calculate the predicted class and store it in Y_{pred}
    calculate the accuracy for s_3
    select the C_{opt} having the best accuracy in s_3
  end for
  Predict the classes for s_4
  Calculate the accuracy
  Output: C_{opt}, α_{opt}, $\theta_{opt}$for all h
end for

erably less. However, in case of 'Heart Disease' and 'Liver Disorder' datasets, the improvement seems to be very high. The accuracies obtained using 'Breast Cancer' dataset with the proposed methodology (when $h = 4$ number of parts are optimally obtained) are better as compared to that of other methods. Considerably higher accuracy is obtained in 'Liver Disorder' data set when $h > 2$ than that obtained when solved with classic SVM, i.e., $h = 1$.

Using performance matrix, we can observe that we have obtained 74 correct predicted "+1" class, and eight objects, which should be "+1", is predicted as "-1". Similarly, nine instances are obtained where the actual class is "-1" but predicted as "+1", and 54 instances are obtained where actual and predicted classes are same, i.e., "-1".

3.2 Datasets used: Our methodology is applied on 'Breast-cancer' Wisconsin [1]. Data type here is multivariate integer type and having 699 number of records with 10 attributes, it has label '2' and '4' and published in year 1992. We used this dataset because using functional nature of the data we can show the performance of the proposed algorithm using this type of dataset.

3.3 Quantitative performance comparison: The maximum accuracy we obtained for "Breast Cancer" using this method is 88.03%. For 'Heart Disease', and 'Liver Disorder' the maximum accuracy obtained is 89.38 % and 90.25%, respectively. In addition, for 'Cryptography', 'Audit', and 'Primary Tumor' datasets, the accuracies obtained are 88.72%, 84.57%, and 89.41%, respectively. The algorithm works efficiently with reportedly less running time. The different classes are easily identified depending on the sub-intervals, and hence, our proposed method creates the separation easier. We observed that the accuracy improvement completely depends on the

[1] `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29`

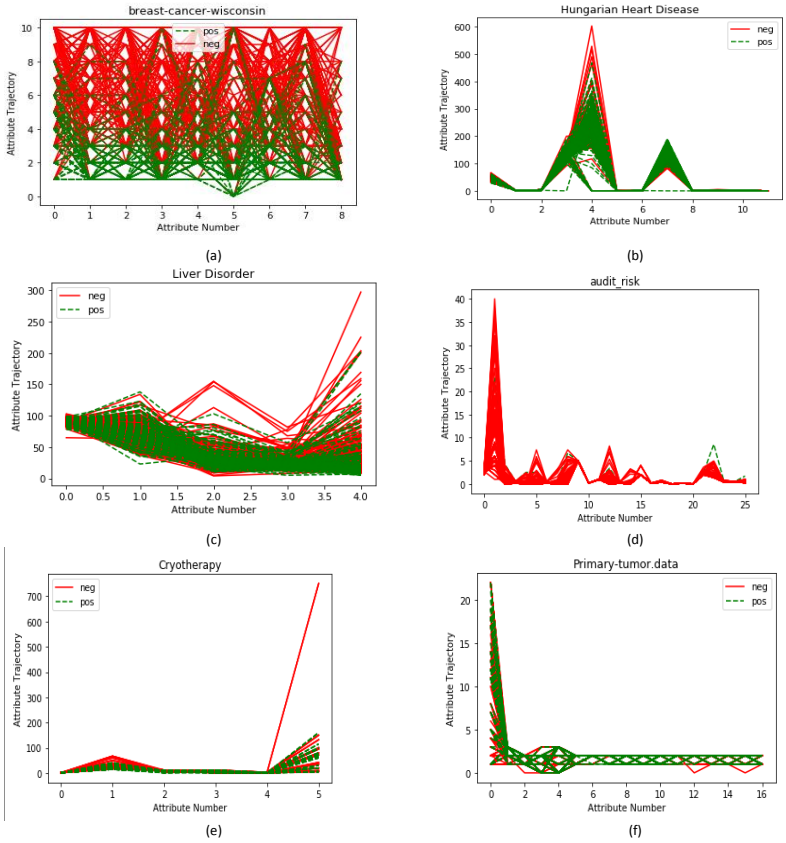

Fig. 1 The results of analyses using (a) Breast Cancer, (b) Heart Disease, (c) Liver Disorder, (d) Audit, (e) Cryptography, and (f) Real datasets.

Table 1 Performance comparison of SVM.

Datasets	**Classic SVM**					
	***k*-NN**	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
Breast Cancer	79.56	82.34	84.15	83.29	88.03	85.35
Heart Disease	75.46	81.47	85.65	85.65	83.22	89.38
Liver Disorder	80.43	86.27	82.64	90.25	87.39	89.45
Audit	82.09	83.29	88.72	86.56	86.56	84.24
Cryptography	76.42	79.81	79.92	82.34	84.57	83.64
Primary Tumor	79.19	83.45	88.47	88.93	86.27	89.41

Table 2 Count of "-1" and "+1" class in the datasets

Datasets	# Records	# Records label (-1)	# Records label (+1)
Breast Cancer	699	338	361
Heart Disease	294	188	106
Liver Disorder	345	150	195
Cryptography	90	42	48
Audit	776	305	471
Primary Tumor	339	224	115

dataset chosen. This method also helps us to find sub-interval of special interest. To get our algorithm work properly, we need to consider many different aspects, such as the count of folds, the number of values C present in the grid chosen, the max value of iterations in the alternate method, and also, the count of runs in multi-start. It is observed that there is no linear increase in the total time required when the code is run in parallel. In this process, we observe that the training takes the maximum amount of time. Once the training is completed, the classification becomes very quick. This is because the only task now left is to calculate the label score shown in Eq. (1) and follow the given rule to classify into different classes. The size of dataset highly influences the resolution of the problems of optimization given in Eq. (2) and Eq.(5). The increase in time is not linear in this method as the code keeps running parallely. Hence, there is reduction in the time taken. In addition, there is increase in running time because of the problems being used are nested here. This is because the optimal solution for simple models is obtained which is then used by the complex optimization problems. However, our method is more efficient than the pre-existing algorithms.

4 Conclusions

In this paper, we presented how the conventional kernel is converted to the dynamic functional bandwidth kernel for FDA. As FD is based on time, it is required to obtain the optimized values of the functional bandwidth kernel. Here, the number of values of functional bandwidth kernel is considered as same as the number of attributes, thereby making the overall process dynamic in nature. Instated of hinge loss, we have used Hill climbing loss, which is more generic, thereby increasing the classification accuracy. As a future study, one may incorporate the fuzzy set theory to solve the optimization problem to obtain more accurate values of functional bandwidth kernel within a less time.

References

1. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing 69(7-9), 730–742 (2006)
2. Martin-Barragan, B., Lillo, R., Romo, J.: Interpretable support vector machines for functional data. European Journal of Operational Research 232(1), 146–155 (2014)

3. Laukaitis, A., Račkauskas, A.: Functional data analysis for clients segmentation tasks. European journal of operational research 163(1), 210–216 (2005)
4. Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., Martín-Barragán, B.: Functional-bandwidth kernel for support vector machine with functional data: An alternating optimization algorithm. European Journal of Operational Research 275(1), 195–207 (2019)
5. Sarkar, S., Patel, A., Madaan, S., Maiti, J.: Prediction of occupational accidents using decision tree approach. In: 2016 IEEE Annual India Conference (INDICON). pp. 1–6. IEEE (2016)
6. Sarkar, S., Pateshwari, V., Maiti, J.: Predictive model for incident occurrences in steel plant in india. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–5. IEEE (2017)
7. Verma, A., Chatterjee, S., Sarkar, S., Maiti, J.: Data-driven mapping between proactive and reactive measures of occupational safety performance. In: Industrial Safety Management, pp. 53–63. Springer (2018)
8. Sarkar, S., Vinay, S., Raj, R., Maiti, J., Mitra, P.: Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research 106, 210–224 (2019)
9. Sarkar, S., Chain, M., Nayak, S., Maiti, J.: Decision support system for prediction of occupational accident: a case study from a steel plant. In: Emerging Technologies in Data Mining and Information Security, pp. 787–796. Springer (2019)
10. Sarkar, S., Raj, R., Vinay, S., Maiti, J., Pratihar, D.K.: An optimization-based decision tree approach for predicting slip-trip-fall accidents at work. Safety science 118, 57–69 (2019)
11. Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., Martín-Barragán, B.: Variable selection in classification for multivariate functional data. Information Sciences 481, 445–462 (2019)
12. Sarkar, S., Pramanik, A., Maiti, J., Reniers, G.: Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. Safety Science 125, 104616 (2020)
13. Sarkar, S., Pramanik, A., Khatedi, N., Maiti, J.: An investigation of the effects of missing data handling using 'r'-packages. In: Data Engineering and Communication Technology, pp. 275–284. Springer (2020)
14. Sarkar, S., Vinay, S., Pateshwari, V., Maiti, J.: Study of optimized svm for incident prediction of a steel plant in india. In: 2016 IEEE Annual India Conference (INDICON). pp. 1–6. IEEE (2016)
15. Sarkar, S., Ejaz, N., Maiti, J.: Application of hybrid clustering technique for pattern extraction of accident at work: a case study of a steel industry. In: 2018 4th International Conference on Recent Advances in Information Technology (RAIT). pp. 1–6. IEEE (2018)
16. Berrendero, J.R., Cuevas, A., Torrecilla, J.L.: Variable selection in functional data classification: a maxima-hunting proposal. Statistica Sinica pp. 619–638 (2016)
17. Kadri, H., Duflos, E., Preux, P., Canu, S., Davy, M.: Nonlinear functional regression: a functional rkhs approach (2010)
18. Xu, Y., Wang, L., Wang, S.y., Liu, M.: An effective teaching–learning-based optimization algorithm for the flexible job-shop scheduling problem with fuzzy processing time. Neurocomputing 148, 260–268 (2015)
19. Colson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Annals of operations research 153(1), 235–256 (2007)