Classification of ECG signals based on functional data analysis and machine learning techniques

Classificazione dei segnali ECG basata sull'analisi dei dati funzionali e tecniche di apprendimento automatico

Mohammed Sabri, Fabrizio Maturo, Rosanna Verde, Jamal Riffi, Ali Yahyaouy and Hamid Tairi

Abstract High-dimensional data classification is always a challenging task due the so-called curse of dimensionality issue. This study proposes a two-steps supervised classification technique for high-dimensional time series treated as functional data. The first phase is based on the idea of extracting additional knowledge from the data using unsupervised classification by means of a new distance that considers the original curves and their derivatives. The second step involves functional supervised classification of the new patterns discovered. Particularly, a Random Forest classifier is built using the new labels obtained in the first step. The experiments on ECG data and comparison with the classical approaches show the effectiveness and exciting improvement in terms of accuracy.

Abstract La classificazione dei dati ad alta dimensionalità è un problema complesso a causa della cosiddetta maledizione della dimensionalità. Questo studio propone una tecnica di classificazione supervisionata in due fasi per serie temporali ad alta dimensionalità trattate come dati funzionali. La prima fase si basa sull'idea di estrarre informazioni dai dati utilizzando una classificazione non supervisionata basata su una nuova distanza. La seconda fase prevede la classificazione supervisionata funzionale considerando i nuovi la-

Mohammed Sabri
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy.
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco, e-mail: sabri.mohammed@unicampania.it

Fabrizio Maturo, Rosanna Verde
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: fabrizio.maturo@unicampania.it, rosanna.verde@unicampania.it

Jamal Riffi, Ali Yahyaouy, Hamid Tairi
Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco e-mail: riffi.jamal@gmail.com, ali.yahyaouy@usmba.ac.ma, htairi@yahoo.fr

bels scoperti. In particolare, viene utilizzato il random forest utilizzando le nuove etichette. Gli esperimenti sui dati ECG e il confronto con i classici approcci di classificazione funzionale mostrano l'efficacia del metodo e un ottimo miglioramento dell'accuratezza del classificatore.

Key words: Functional Data Analysis (FDA), supervised classification, Functional K-means, Functional Random Forest

1 Introduction

The main objective of the electrocardiogram (ECG) analysis is to detect the life-threatening arrhythmias accurately for appropriate treatment in order to save life. During the last decades, several methods were reported for automatic ECG beat classifications. ECG data is a classic example of high-dimensional data, and therefore their classification requires adequate statistical techniques. Classification of high-dimensional data is a fast-growing research area, driven by a need for methods to deal with the increasing availability of data coming from sensors and biomedical devices. However, due to the course of dimensionality problem, it is always challenging to build an accurate model that is reasonably flexible and yet feasible to fit.

A possible approach to deal with high-dimensional data is functional data analysis (FDA), i.e. considering each time series as a single entity given by a functional object [1]. In this context, recently, several metrics and semi-metrics have been proposed to compute the similarity among curves, and many approaches have been suggested to deal with functional data classification problems.

Our starting idea to improve the classical functional classifiers' performance is to consider ECG data as functional data in the time domain [3, 4] and propose an original two-phase classification approach. In the first step, a functional clustering algorithm is used to discover new patterns in the original classes, e.g. the functional K-means algorithm. Then, supervised classification based on Random Forest (RF) is applied exploiting the additional information on the new labels coming from the first step. The basic idea is to get additional knowledge in the training data to improve the power of the final classifier. Also, we define a novel distance used to measure the similarity among functional samples by considering also the information of their derivatives.

2 The two-phases functional classification procedure

FDA analyses samples where each observation arises from a function varying over a continuum. For ECG data, the continuum is function. Thus, we consider a set X of ECG signals where y_i $(i = 1, \ldots, N)$ indicate the class label of the $i-th$ signal. The basic idea is to represent each ECG signal as a functional data that can be expressed as a linear combination of basis functions, e.g. b-splines. B-spline. Assuming fixed basis, each i-th signal can be approximated as follows:

$$x_i(t) = \sum_{j=1}^{p} c_{i,j}\psi_j(t) \tag{1}$$

where ψ_j are p known basis functions and $c_{i,j}$ are the corresponding coefficients to be estimated. As the ECG data is usually a non-periodic data, B-spline basis system are used. The B-splines basis coefficients are estimated by the ordinary least squares method, minimizing the sum of squared residuals [1, 2].
The first advantage of using FDA is that we consider the whole shape of the curves to compute the similarity among functional data and classify them. The second advantage is that we can deal with the curse of dimensionality issue by using a low number of coefficients via dimensionality reduction techniques able to create independent features.
The main idea of our proposal is using clustering to discover new patterns in the dataset, and thus exploiting a combination of supervised and unsupervised methods to improve the performance of a functional classifier.
In addition, this research proposes a new distance to measure the similarity between two functional data. The latter semi-metric also involves the use of derivatives to consider the similarity between curves to take into account additional behaviours of the functions.
Given two functional data $x_i(t)$ and $x_j(t)$ from a data set X, a new similarity metric between $x_i(t)$ and $x_j(t)$ is defined as

$$d^2(x_i(t), x_j(t)) = d_{ij}^{(0)} + d_{ij}^{(1)} + d_{ij}^{(2)} \tag{2}$$

with
$d_{ij}^{(0)} = \frac{1}{\int_T \sigma_{x(t)}(dt)} \int_T (x_i(t) - x_j(t))^2 dt$
$d_{ij}^{(1)} = \frac{1}{\int_T \sigma_{Dx(t)}(dt)} \int_T (Dx_i(t) - Dx_j(t))^2 dt$
$d_{ij}^{(2)} = \frac{1}{\int_T \sigma_{D^2x(t)}(dt)} \int_T (D^2x_i(t) - D^2x_j(t))^2 dt$
and where $Dx_i(t)$ is the first-order functional derivative of the i-th curve, and $D^2x_i(t)$ is the second-order functional derivative of the i-th curve.

Silhouette analysis [8] is used to determine the most suitable number of subgroups of each original label by using the new distance defined in Equation (2); the functional k-means clustering algorithm, based on the distance defined in Equation (2), is implemented to discover the functional elements of each subgroup in the training set by employing the number of clusters we get from the silhouette method. The entire process of the functional k-means clustering algorithm based on the distance d in Equation (2) is described as in Algorithm 1.

Algorithm 1 The functional k-means clustering algorithm based on the new metric

input:
- X_g group of curves in the group g ($g = 1, ..., G$)
- K_g number of clusters (subgroups) found in the group g

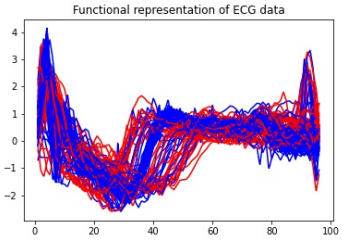
Output: - $labels_g$: the new labels of the group X_g

for g $\leftarrow$ 1 to G **do**
  Randomly choose K_g samples as the initial centroids $\mu_1(t), \ldots, \mu_{K_g}(t)$
  while stopping criterion has not been met **do**
    for i $\leftarrow$ 1 to K_g **do**
      for $x(t) \in X_g$ **do**
        $j \leftarrow \arg\min_i \; d(\mu_i(t), x(t))$
        $G_j \leftarrow G_j \bigcup x(t)$
      end for
      for m $\leftarrow$ 1 to K_g **do** $\mu_m(t) \leftarrow \dfrac{1}{|G_m|} \displaystyle\sum_{x(t) \in G_m} x(t)$
      end for
    end for
  end while
  $labels_g < -G$
end for

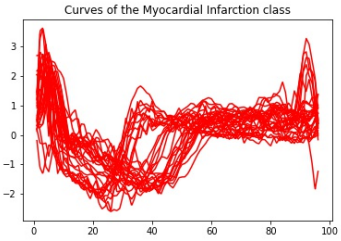
Afterwards, in the second phase, the B-spline coefficients are used as features in the input of the RF algorithm to train and validate the functional classification model.

3 An application using ECG data with a binary outcome

We apply the method described in previous section to the well-known ECG dataset [7] formatted by R. Olszewski as part of his thesis "Generalized feature extraction for structural pattern recognition in time-series data" at Carnegie Mellon University, 2001. Our objective is to create a functional model to classify ECG curves "normal" or "myocardial infarction".
Figure 1 illustrates the smoothed versions of the original signals computed using Equation (1).



(a) Functional representation of the ECG data

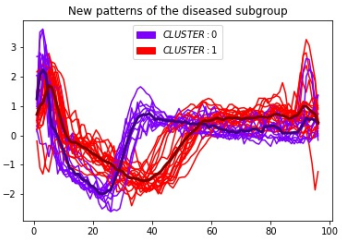


(b) Myocardial Infarction curves in the training set

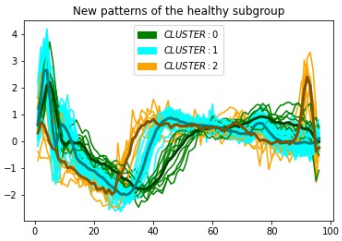
(c) Healthy curves in the training set

Fig. 1: Smoothed ECG curves in the training set.

After using the silhouette analysis with the newly defined metric for the normal and myocardial infarction classes, the most suitable number of clusters for the myocardial infarction original class is two whereas, for healthy people, we have three subgroups. Based on the latter results, the functional K-means is applied. Figure 2 shows the new patterns discovered for each subgroup in terms of functional subsets.



(a) New subgroups discovered in diseased subgroup



(b) New subgroups discovered in healthy subgroup

Fig. 2: New patterns (subgroups) discovered in the original groups of the training data.

Table 1 shows the effectiveness of the proposed two-steps functional classification approach based on the novel distance. A maximum accuracy of 87%

is obtained using RF and K-means with the new distance. Instead, using a classical supervised classification, without the two-steps procedure, we get an accuracy of 80%.

Table 1: Model performance, the combinations of Random Forest and K-means

	Random Forest			
	Without K-means	Classical K-means	K-means with the Euclidean functional distance	K-means with the new distance
Accuracy %	80	83	85	87
F1-Score	79	82	85	87
Specificity	69	64	75	83

References

1. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. 2nd Ed., Springer, New York (2005)
2. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York (2006)
3. Zhou, Y., Sedransk, N.: Functional data analytic approach of modeling ECG T-wave shape to measure cardiovascular behavior. The Annals of Applied Statistics **3**(4), 1382-1402 (2009)
4. Jacques, J. and Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis, **71** 92–106 (2014)
5. Möller, A., Tutz, G., and Gertheiss, J.: Random forests for functional covariates. Journal of Chemometrics **30**(12), 715 – 725 (2016)
6. Huang, Q., Li, Y. and Liu, P.:Short term load forecasting based on wavelet decomposition and random forest. Proceedings of the Workshop on Smart Internet of Things. ACM. p. 2 (2017)
7. Olszewski R. Generalized feature extraction for structural pattern recognition in time-series data. Ph.d. dissertation, School of Computer Science, Carnegie Mellon University, (2001)
8. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)