COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 108 (2012) 250–261

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Unsupervised feature relevance analysis applied to improve ECG heartbeat clustering

J.L. Rodríguez-Sotelo[a], *D. Peluffo-Ordoñez*[b], *D. Cuesta-Frau*[c,*], *G. Castellanos-Domínguez*[b]

[a] *Universidad Autónoma de Manizales, Grupo Automática, D. Ing. Electrónica y Automatización, Antigua estación de ferrocarril, Manizales, Colombia*

[b] *Universidad Nacional de Colombia, D. Ing. Eléctrica, Electrónica y Computación, Km. 9, Vía al aeropuerto, Campus la Nubia, Manizales, Colombia*

[c] *Technological Institute of Informatics, Polytechnic University of Valencia, Campus Alcoi, Plaza Ferrándiz y Carbonell, 2, 03801 Alcoi, Spain*

ARTICLE INFO

Article history:
Received 30 August 2011
Received in revised form
3 April 2012
Accepted 17 April 2012

Keywords:
Electrocardiogram analysis
Feature selection
Relevance analysis
Heartbeat classification
$Q-\alpha$ algorithm

ABSTRACT

The computer-assisted analysis of biomedical records has become an essential tool in clinical settings. However, current devices provide a growing amount of data that often exceeds the processing capacity of normal computers. As this amount of information rises, new demands for more efficient data extracting methods appear.

This paper addresses the task of data mining in physiological records using a feature selection scheme. An unsupervised method based on relevance analysis is described. This scheme uses a least-squares optimization of the input feature matrix in a single iteration. The output of the algorithm is a feature weighting vector.

The performance of the method was assessed using a heartbeat clustering test on real ECG records. The quantitative cluster validity measures yielded a correctly classified heartbeat rate of 98.69% (specificity), 85.88% (sensitivity) and 95.04% (general clustering performance), which is even higher than the performance achieved by other similar ECG clustering studies. The number of features was reduced on average from 100 to 18, and the temporal cost was a 43% lower than in previous ECG clustering schemes.

1. Introduction

The computer-assisted analysis of biomedical records has become an essential tool in clinical settings. The widespread access to portable medical devices or new personal devices such as cell phones, smartphones, pdas, tablets, and wearable devices, is boosting the amount of biomedical data available. These devices provide a growing amount of data that often exceeds the processing capacity of affordable computers. As this amount of biosignal data rises, new demands for more efficient information extracting methods appear [1].

A number of algorithms have been proposed for knowledge discovery and management in medical databases [2]. These methods are aimed at turning the huge amount of information in these databases into a more manageable source. This objective can be achieved by removing useless data such as noise or outliers [3], selecting only a data subset [4], performing a data mining process to capture data patterns or relationships [5], or by obtaining an efficient lower dimension representation of the data [6].

Feature selection algorithms are dimensionality reduction methods often associated to data mining tasks of classificationor clustering [1]. These methods provide a reduced set

* *Corresponding author*. Tel.: +34 651028565.
E-mail addresses: jlrodriguez@autonoma.edu.co (J.L. Rodríguez-Sotelo), dcuesta@disca.upv.es (D. Cuesta-Frau), cgcastellanosd@unal.edu.co (G. Castellanos-Domínguez).

0169-2607/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2012.04.007

of the input features while preserving the relevant discriminatory information. Feature matrix projection methods and its derivatives, specially principal component analysis (PCA) based methods, are probably the most popular of such relevance analysis schemes. PCA related methods have been applied successfully to many biomedical signals: ECG [7–9], EEG [10,11], RR series [12], respiration [13]; and in diagnosis applications related to the Alzheimer's disease [14], Parkinson's disease [15], and many more such as in [16,17], or [18], among others.

We describe in this paper a new feature matrix projection scheme for unsupervised relevance analysis. This method can be applied to any biomedical signal where a clustering process is involved. It outperforms the conventional PCA methods in terms of generalization, complexity, and feature discriminatory information. The strong asymmetry frequently found in biomedical data sometimes prevent PCA methods from being used, since they assume a uniform data class distribution [19]. In addition, PCA methods are also very sensitive to the presence of noise and outliers [20].

The algorithm proposed is based on a least-squares optimization scheme of the feature matrix that iteratively converges to a local maximum of a relevance function [21]. The output of the method is a feature significance score that enables the exclusion of superfluous features and a weighted representation of the remaining ones. We have further improved the original method by reducing the number of iterations to 1. Consequently, the computational requirements to obtain the final feature set is lower, whereas its information content remains almost unchanged.

The performance of the method presented was assessed using a heartbeat clustering test on ECG records. ECGs are one of the most used biomedical signals, and long-term records, where data mining might be a powerful tool, are quite common [4]. Labelled records from the MIT/BIH database [22], containing different types of arrhythmias, were used in the experiments.

A general feature extraction procedure was applied to the segmented heartbeats. The feature set went then through the relevance analysis stage proposed before entering the clustering process. The objective of the experiments was to assess the quality of the heartbeat partition obtained using this method of relevance analysis, both in terms of accuracy and temporal cost.

2. Method

The objective of a feature relevance analysis is to find a simpler representation of the input features that preserves most of their significant information according to a given criterion function [23]. A tradeoff must be found between the quantity and/or quality of the input features, and the separability of the objects from which they were extracted. Feature selection has many important potential benefits that makes it suitable in a number of clustering applications [24].

Our method employs a feature relevance analysis scheme to filter the input data, and preserve and score only those features that maximize the information content. Although this method can be used in any case where a clustering process of a feature vector set takes place, we focused our attention on ECG data. ECG records are one of the most common biomedical signals, and high-resolution and/or long-term ECGs are probably the best candidates for data mining approaches [4].

2.1. *Heartbeat feature selection algorithm*

The aim of the method is to compute weighting feature values that enable an effective data dimensionality reduction along with a proper feature relevance ranking. To accomplish this goal, a distance-based least-squares optimization scheme is proposed.

2.1.1. *Definitions*

Let $\mathbf{x}_j \in \mathbb{R}^p$ be a heartbeat feature vector, $1 \le j \le n$. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a matrix that results from arranging all $\mathbf{x}_j$ vectors in rows. Let $\mathbf{Y} \in \mathbb{R}^{n\times p}$ be a linear projection of $\mathbf{X}$ given by $\mathbf{Y} = \mathbf{X}\mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^{p\times p}$ is an unknown orthogonal projection matrix. If the number of features is reduced from p to q, we have to introduce the truncated version of the previous matrices, namely $\hat{\mathbf{V}} \in \mathbb{R}^{p\times q}$, $\hat{\mathbf{Y}} \in \mathbb{R}^{n\times q}$, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{V}}$, and the least squares estimated matrix $\hat{\mathbf{X}} \in \mathbb{R}^{n\times p}$, $\hat{\mathbf{X}} = \hat{\mathbf{Y}}\hat{\mathbf{V}}^\top$.

Finally, let $\boldsymbol{\alpha} \in \mathbb{R}^p$ be an unknown feature weighting vector that can be arranged in an affinity symmetric positive definite matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{X}\mathrm{diag}(\boldsymbol{\alpha})\mathbf{X}^\top$, with $||\boldsymbol{\alpha}|| = 1$.

2.1.2. *Mathematical background*

The dissimilarity between the original feature matrix $\mathbf{X}$ and the reconstructed version $\hat{\mathbf{X}}$ can be quantified by the euclidean norm, that is, $||\mathbf{X} - \hat{\mathbf{X}}||$. Since this dissimilarity has to be minimized in terms of $\hat{\mathbf{V}}$, and $\mathbf{A}$, the initial objective function can be expressed as:

$$\min_{\hat{\mathbf{V}}} ||\mathbf{X} - \hat{\mathbf{X}}||_{\mathbf{A}}^2 = \min_{\hat{\mathbf{V}}} ||\mathbf{X} - \hat{\mathbf{Y}}\hat{\mathbf{V}}^\top||_{\mathbf{A}}^2 \tag{1}$$

where we use a squared version for simplicity. Applying the inner product definition and trace properties to Eq. (1), it can be rewritten as:

$$\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X}^\top - \hat{\mathbf{X}}^\top\mathbf{A}\mathbf{X}^\top - \mathbf{X}^\top\mathbf{A}\hat{\mathbf{X}}^\top + \hat{\mathbf{X}}^\top\mathbf{A}\hat{\mathbf{X}}^\top) \tag{2}$$

Another term of interest is $||\mathbf{X}||_{\mathbf{A}}^2$, given by:

$$||\mathbf{X}||_{\mathbf{A}}^2 = \mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X}) = \mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{X}^\top) = \sum_{i=1}^{p} \gamma_i\lambda_i \tag{3}$$

where λ_i and $\mathbf{v}_i$ denote respectively the i-th eigenvalue and eigenvector of $\mathbf{X}^\top\mathbf{X}$ and $\gamma_i = \lambda_i^{-1}\mathbf{v}_i^\top\mathbf{X}^\top\mathbf{A}\mathbf{X}\mathbf{v}_i$. Arranging Eq. (2) according to formulation for Eq. (3), it results in the following new expression:

$$\sum_{i=1}^{p} \gamma_i\lambda_i - 2\sum_{i=1}^{q} \gamma_i\lambda_i + \sum_{i=1}^{q} \gamma_i\lambda_i = \sum_{i=q+1}^{p} \gamma_i\lambda_i \tag{4}$$