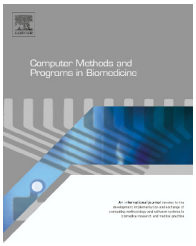
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 238–246

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



Multichannel biomedical time series clustering via hierarchical probabilistic latent semantic analysis

Jin Wang[a,b,*], *Xiangping Sun*[b,c], *Saeid Nahavandi*[b], *Abbas Kouzani*[c], *Yuchuan Wu*[d], *Mary She*[b,c]

[a] *School of Computer Science & Software Engineering, The University of Western Australia, Australia*
[b] *Center for Intelligent Systems Research, Deakin University, Australia*
[c] *School of Engineering, Deakin University, Australia*
[d] *School of Mechanical Engineering and Automation, Wuhan Textile University, Wuhan, PR China*

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form
20 June 2014
Accepted 22 June 2014

Keywords:
Bag-of-words
PLSA
Topic model
Unsupervised learning
ECG

ABSTRACT

Biomedical time series clustering that automatically groups a collection of time series according to their internal similarity is of importance for medical record management and inspection such as bio-signals archiving and retrieval. In this paper, a novel framework that automatically groups a set of unlabelled multichannel biomedical time series according to their internal structural similarity is proposed. Specifically, we treat a multichannel biomedical time series as a document and extract local segments from the time series as words. We extend a topic model, i.e., the Hierarchical probabilistic Latent Semantic Analysis (H-pLSA), which was originally developed for visual motion analysis to cluster a set of unlabelled multichannel time series. The H-pLSA models each channel of the multichannel time series using a local pLSA in the first layer. The topics learned in the local pLSA are then fed to a global pLSA in the second layer to discover the categories of multichannel time series. Experiments on a dataset extracted from multichannel Electrocardiography (ECG) signals demonstrate that the proposed method performs better than previous state-of-the-art approaches and is relatively robust to the variations of parameters including length of local segments and dictionary size. Although the experimental evaluation used the multichannel ECG signals in a biometric scenario, the proposed algorithm is a universal framework for multichannel biomedical time series clustering according to their structural similarity, which has many applications in biomedical time series management.

1. Introduction

With the development of modern recording technology and reduction of hardware cost, more and more biomedical time series such as Electrocardiography (ECG) signals are recorded to monitor human physiological condition. How to effectively and efficiently manage and analyse a large amount of physiological signals is a big challenge. Traditionally, these physiological signals are manually managed and analysed by

* *Corresponding author at*: School of Computer Science & Software Engineering, The University of Western Australia, Australia. Tel.: +61 406417698.
E-mail addresses: jay.wangjin@gmail.com, wjin@deakin.edu.au (J. Wang) .

http://dx.doi.org/10.1016/j.cmpb.2014.06.014
0169-2607/

medical experts. However, manual management and inspection are time-consuming and labour-intensive. Even worse, false hit rates by operators may increase considerably for a long-term inspection and management, since it is difficult for human to keep a high level of concentration for a long time. Therefore, automatic methods that can help medical exporters effectively manage and inspect a large amount of physiological time series are very valuable.

One of the automatic methods for biomedical time series management and inspection is time series clustering [1–5], which groups a collection of time series with no prior label information according to their internal structural similarity. The time series clustering makes biomedical time series management such as bio-signals archiving and retrieval much easier. For biomedical time series clustering, it is of importance to extract discriminative features to characterize the time series. Some state-of-the-art works extract features from time domain [6–8] while some others transform the time series into frequency domain [9–11]. However, most of these methods are limited to extract internal structural similarity information. To this end, Wang et al. [1,12] and Lin and Li [13] proposed a bag-of-words/patterns representation that was originally developed for text document analysis to effectively capture the structural similarity information of time series. In the bag-of-words/patterns representation [1,12,13], time series are treated as documents and local segments are extract from the time series as words. A time series is then represented as a histogram of the number of codewords occurred in the time series.

Based on the bag-of-words/patterns representation, probabilistic topic models such as probabilistic Latent Semantic Analysis (pLSA) [14] and Latent Dirichlet Allocation (LDA) [15] were extended to cluster a set of unlabelled biomedical time series according to their structural similarity [1,16]. It is demonstrated that the probabilistic topic model is able to naturally model the generative process of the words/patterns in time series, which provides very promising clustering performance [1,16].

However, the clustering framework proposed in [1,16] was developed for single-channel time series analysis. In real clinical applications, many biomedical signals are recorded in multiple channels. For instance, ECG signals are always recorded in more than one channel to provide more comprehensive clinical information. In this paper, we extend the bag-of-words representation and the probabilistic topic models for multichannel time series analysis. Similar to the bag-of-words representation in single-channel time series analysis [12], we treat a multichannel time series as a document and extract local segments from each channel of the time series as words. Based on the bag-of-words representation, we extended the topic models to analyse multichannel biomedical time series in an unsupervised manner. Specifically, a hierarchical pLSA (H-pLSA) [17] that was originally developed for visual motion analysis is extended to cluster multichannel biomedical time series.

The hierarchical pLSA developed in [17] models visual motion using a two-layer pLSA. Local motion behaviours are modelled in the first layer, and global motion behaviours are discovered by the global pLSA. In this paper, we extend the hierarchical pLSA to automatically discover categories of multichannel biomedical time series. In the first layer, we model each channel of the time series using a local pLSA model. The local topics extracted from each channel of the time series are then treated as words in the second-layer pLSA, i.e., global pLSA. The categories of the multichannel time series are automatically discovered by the global pLSA.

The main contribution of the paper is 3-fold: (i) the bag-of-words model was extended to represent multichannel time series; (ii) based on the bag-of-words representation, a hierarchical pLSA (H-pLSA) was developed for multichannel time series clustering; (iii) a series of experiments were conducted to investigate the effectiveness and robustness of the H-pLSA for multichannel time series clustering.

The rest of this paper is organized as follows. In Section 2, we introduce how to construct a bag-of-words representation for multichannel time series. The details of the hierarchical pLSA are illustrated in Section 3. The experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2. Bag-of-words representation

The method in [12] continuously slides a pre-defined length window along a single-channel time series to extract local segments, and constructs a bag-of-words representation for single-channel time series analysis. Similarly, for multichannel time series, we continuously slide a window with pre-defined length along each channel of a time series to extract a group of segments. Each segment is then ℓ_2 normalized to be a feature vector, i.e., each of the feature vectors is normalized to be a ell_2-unit.

Similar to the dictionary construction in [12], we cluster all the local segments extracted from all the channels of time series by k-means clustering to construct the dictionary, which contains a set of codewords (i.e., cluster centres estimated by the k-means clustering). Denoting the normalized local segments extracted from time series as: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I] \in \mathbb{R}^{l\times I}$, the dictionary construction by k-means clustering is formulated as an optimization problem:

$$\begin{aligned} &\min_{\mathbf{D}\in\mathbb{R}^{l\times N},\mathbf{V}\in\mathbb{R}^{K\times I}} \sum_{i=1}^{I} \|\mathbf{x}_i - \mathbf{D}\mathbf{v}_i\|_2, \\ &\text{s.t.} \quad card(\mathbf{v}_i) = 1, |\mathbf{v}_i| = 1, \forall i, \mathbf{v}_i \geq 0, \end{aligned} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{l\times N}$ is the learned dictionary, i.e, clustering centers. The unit-basis vector $\mathbf{v}_i$ indicates the clustering index of the local segment $\mathbf{x}_i$. The constraint means that the vector $\mathbf{v}_i$ only has one component ($card(\mathbf{v}_i) = 1$) that equals to one ($|\mathbf{v}_i| = 1$) and all the other components are zero.

Denoting the learned dictionary as $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N] \in \mathbb{R}^{l\times N}$, and the ith segment from the hth channel as x_i^h, the segment x_i^h is assigned the codeword that is nearest, i.e., $c^* = \arg\min_j dist(\mathbf{d}_j, \mathbf{x}_i^h)$, where $dist(\cdot, \cdot)$ denotes the Euclidean distance function. It is worth noting that the dictionary is universal for all the multichannel time series and only needs to be learned for once.