Multichannel biomedical time series clustering via hierarchical probabilistic latent semantic analysis

Jin Wang\textsuperscript{a,b,}\textsuperscript{*}, Xiangping Sun\textsuperscript{b,c}, Saeid Nahavandi\textsuperscript{b}, Abbas Kouzani\textsuperscript{c}, Yuchuan Wu\textsuperscript{d}, Mary She\textsuperscript{b,c}

\textsuperscript{a} School of Computer Science & Software Engineering, The University of Western Australia, Australia
\textsuperscript{b} Center for Intelligent Systems Research, Deakin University, Australia
\textsuperscript{c} School of Engineering, Deakin University, Australia
\textsuperscript{d} School of Mechanical Engineering and Automation, Wuhan Textile University, Wuhan, P.R. China

\textbf{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.

© 2014 Elsevier Ireland Ltd. All rights reserved.

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/© 2014 Elsevier Ireland Ltd. All rights reserved.
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 used to construct 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 $X = [x_1, x_2, \ldots, x_l] \in \mathbb{R}^{l \times d}$, the dictionary construction by $k$-means clustering is formulated as an optimization problem:

$$
\min_{D \in \mathbb{R}^{k \times N}, V \in \mathbb{R}^{k \times d}} \sum_{l=1}^{l} \|x_l - Dv_l\|_2^2.
$$

s.t. $\text{card}(v_i) = 1$, $|v_i| = 1$, $\forall i$, $v_i \geq 0$.

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

Denoting the learned dictionary as $D = [d_1, d_2, \ldots, d_N] \in \mathbb{R}^{k \times N}$, and the $ith$ segment from the $ith$ channel as $x_i^t$, the segment $x_i^t$ is assigned the codeword that is nearest, i.e., $c^* = \arg \min_i \text{dist}(d_i, x_i^t)$, where $\text{dist}(\cdot, \cdot)$ 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.

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 used to construct 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 $X = [x_1, x_2, \ldots, x_l] \in \mathbb{R}^{l \times d}$, the dictionary construction by $k$-means clustering is formulated as an optimization problem:

$$
\min_{D \in \mathbb{R}^{k \times N}, V \in \mathbb{R}^{k \times d}} \sum_{l=1}^{l} \|x_l - Dv_l\|_2^2.
$$

s.t. $\text{card}(v_i) = 1$, $|v_i| = 1$, $\forall i$, $v_i \geq 0$.

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

Denoting the learned dictionary as $D = [d_1, d_2, \ldots, d_N] \in \mathbb{R}^{k \times N}$, and the $ith$ segment from the $ith$ channel as $x_i^t$, the segment $x_i^t$ is assigned the codeword that is nearest, i.e., $c^* = \arg \min_i \text{dist}(d_i, x_i^t)$, where $\text{dist}(\cdot, \cdot)$ 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.
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات

ISIArticles
مرجع مقالات تخصصی ایران