Neurocomputing 103 (2013) 93–103

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Unsupervised mining of long time series based on latent topic model

Jin Wang [a,c,*], Xiangping Sun [a], Mary F.H. She [a], Abbas Kouzani [b], Saeid Nahavandi [c]

[a] Institute for Technology Research and Innovation, Deakin University, Geelong, VIC 3217, Australia
[b] School of Engineering, Deakin University, Geelong, VIC 3217, Australia
[c] Center for Intelligent Systems Research, Deakin University, Geelong, VIC 3217, Australia

ARTICLE INFO

Article history:
Received 24 October 2011
Received in revised form
14 May 2012
Accepted 16 September 2012
Communicated by S. Choi
Available online 23 October 2012

Keywords:
ECG signals
Bag-of-patterns
Unsupervised learning
pLSA
LDA

ABSTRACT

This paper presents a novel unsupervised method for mining time series based on two generative topic models, i.e., probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). The proposed method treats each time series as a text document, and extracts a set of local patterns from the sequence as words by sliding a short temporal window along the sequence. Motivated by the success of latent topic models in text document analysis, latent topic models are extended to find the underlying structure of time series in an unsupervised manner. The clusters or categories of unlabeled time series are automatically discovered by the latent topic models using bag-of-patterns representation. The proposed method was experimentally validated using two sets of time series data extracted from a public Electrocardiography (ECG) database through comparison with the baseline *k*-means and the Normalized Cuts approaches. In addition, the impact of the bag-of-patterns' parameters was investigated. Experimental results demonstrate that the proposed unsupervised method not only outperforms the baseline *k*-means and the Normalized Cuts in learning semantic categories of the unlabeled time series, but also is relatively stable with respect to the bag-of-patterns' parameters. To the best of our knowledge, this work is the first attempt to explore latent topic models for unsupervised mining of time series data.

1. Introduction

Time series mining has attracted great interest in many communities due to its wide applications in medicine, finance, aerospace and several other industries [1–3]. There exist two fundamental research topics associated with time series mining. The first one is feature representation for time series, and the other one is prediction and classification methods for exploiting the similarity of time series. Some preliminary methods [4,5] directly describe time series in time domain while some others extract features from a transformed domain [6,7]. Most of the previous methods [8,9] are developed for analyzing short sequences or sequences with periodic wave forms. They may be insufficient to represent overall characteristics of long time series that have repetitive but unperiodic wave forms, for instance, long-term Electrocardiography (ECG) and Electroencephalography (EEG) signals. Therefore, it is more appropriate to model sequence-level structural similarity to capture higher-level information of these long time series [10].

Latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) [11] and the Latent Dirichlet Allocation (LDA) [12] that were originally developed for text document analysis provide a statistical approach to semantically summarize and analyze large scale document collections. In this work, motivated by the success of latent topic models in text document analysis, we extend the latent topic models to semantically learn the underlying structure of long time series based on the bag-of-patterns representation [10,13]. Unlike general clustering approaches such as the Normalized Cuts [14] and the *k*-means methods which treat the bag-of-patterns representation as general feature vectors, the topic model naturally models the generative process of local patterns in time series [11,12]. Although the temporal order information of local patterns is ignored in the bag-of-patterns representation, sequence level information is well captured by the topic model. Moreover, the topic model based method has the potential to be used for time series segmentation based on the posterior distribution of topics in a time series.

1.1. Overview of the proposed approach

The paradigm of the proposed unsupervised approach consists of extracting a set of subsequences from each time series, converting each subsequence to a local pattern, and unsupervised learning by the topic models, as shown in Fig. 1.

We slide a window along each time series to extract a set of subsequences with defined length. Then, each subsequence is normalized to have zero mean and standard deviation. After that,

* Corresponding author at: Institute for Technology Research and Innovation, Deakin University, Geelong, VIC 3217, Australia. Tel.: +61 430479874.
E-mail address: jay.wangjin@gmail.com (J. Wang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.09.008

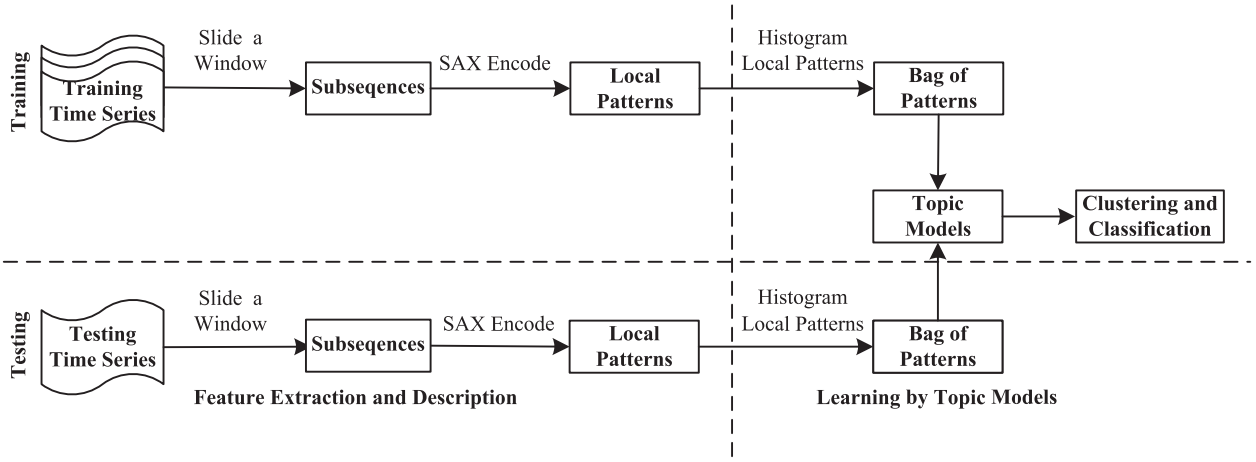

Fig. 1. The framework of the proposed approach for unsupervised time series mining based on topic models.

each normalized subsequence with the same length is converted to a Symbolic Aggregate approXimation (SAX) [15] string, referred to as a local pattern. As a result, a group of local patterns are extracted from each time series that corresponds to a subsequence. We ignore the temporal order of local patterns, and calculate the histogram of the SAX strings in each time series to construct the bag-of-patterns representation.

The two topic models, i.e., the pLSA and the LDA, are employed to learn the probability distributions of local patterns in the time series. Similar to text document analysis, each local pattern is regarded as a word and each time series as a document. The intermediate topics learned by the topic models correspond to the categories of time series. Specifically, we regard each topic in the topic models as one particular category of time series, and set the number of topics as the number of categories. Given a set of new unlabeled time series, the topic models can automatically discover the underlined categories (semantic clusters) through the distribution of topics in the time series.

1.2. Contribution and organization

The main contribution of the paper is twofold. First of all, we present two generative models for mining long time series in an unsupervised manner based on the bag-of-patterns representation. To the best of our knowledge, it is the first time to apply topic models to time series mining. Furthermore, we experimentally compare our method with the baseline *k*-means clustering and Normalized Cuts [14] on two ECG datasets. Experimental results show that our method significantly outperforms the *k*-means and the Normalized Cuts methods.

The paper is organized as follows. In Section 2, we briefly review the related works on unsupervised mining of time series, and the topic models used for clustering and unsupervised classification. Section 3 describes the bag-of-patterns representation for long time series. Section 4 gives the details of the two topic models, i.e., pLSA and LDA. Experimental results on two ECG datasets are analyzed in Section 5. Discussion and conclusion are given in Sections 6 and 7, respectively.

2. Related work

2.1. Unsupervised time series mining

Many feature representations and distance measurement methods have been proposed to mine time series, including Discrete Wavelet Transform (DWT) [16], Symbolic Aggregate approXimation (SAX) [15], Compression-based Dissimilarity Measure (CDM) [17], Dynamic Time Warping (DWT) [18] and so on. The SAX proposed by Lin et al. [15] converted a time series into a string of symbols with lower dimension, and it was later extended to a multiresolution representation in [19]. More recently, Lin and Li [13] proposed an efficient bag-of-patterns representation based on the SAX for encoding long time series. Our approach is based on the bag-of-patterns representation, which considers both local structures and global structures in time series.

There are mainly three groups of clustering methods of time series, i.e., raw-data-based, feature-based and model-based [8]. The raw-data-based methods and the feature-based methods usually use raw data or extracted features with traditional clustering algorithms such as *k*-means and fuzzy *c*-means [20]. By contrast, the model-based approaches assume that the time series data are generated by some kind of models or probability distributions. Our approach belongs to the model-based methods, as the topic models are adopted to model the generative process of the long time series.

2.2. Topic models for clustering and classification

Topic models such as the pLSA and the LDA are originally proposed for text document analysis based on the bag-of-words representation, where a document is characterized by a histogram of word occurrence. Recently, they are extended for object categorization [21,22], action recognition [23,24] and high-level event understanding [25,26] in images and videos.

Sivic et al. [21] adopted the pLSA model to simultaneously discover objects' categories and their spatial layout in images. They treated small local patches extracted from images as "words" and let the intermediate topics directly correspond to the categories of objects. In a similar way, Niebles et al. [23] treated 3-D local patches in videos as "words" and applied two topic models to learn human action categories in an unsupervised manner. The methods in [21–23] extracted local patches from images or videos, and then clustered all the local patches to construct a codebook (dictionary). In contrast, works in [25,26] defined "words" based on motions of pixels in divided pixel-cells. High-level interactions between objects are learned through the probability distribution of intermediate topics. Besides applications in computer vision field, topic models are also widely applied in other areas such as program debugging [27], protein function inferring [28] and chemogenomic profiling [29].