Neurocomputing 149 (2015) 777–787

Contents lists available at ScienceDirect

Neurocomputing

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

Early classification on multivariate time series

Guoliang He*, Yong Duan, Rong Peng, Xiaoyuan Jing, Tieyun Qian, Lingling Wang

State Key Lab of Software Engineering, College of Computer Science, Wuhan University, Wuhan, China

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form
15 July 2014
Accepted 29 July 2014
Communicated by P. Zhang
Available online 13 August 2014

Keywords:
Multivariate time series
Early classification
Feature selection

ABSTRACT

Multivariate time series (MTS) classification is an important topic in time series data mining, and has attracted great interest in recent years. However, early classification on MTS data largely remains a challenging problem. To address this problem without sacrificing the classification performance, we focus on discovering hidden knowledge from the data for early classification in an explainable way. At first, we introduce a method MCFEC (Mining Core Feature for Early Classification) to obtain distinctive and early shapelets as core features of each variable independently. Then, two methods are introduced for early classification on MTS based on core features. Experimental results on both synthetic and real-world datasets clearly show that our proposed methods can achieve effective early classification on MTS.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Multivariate time series (MTS for short) are widely used in many areas such as speech recognition, anomaly detection of EEG/ECG data, science and engineering. MTS data are complex because a MTS sample contains multiple observations at a timestamp. Over a period of time, the collected data of each variable, called a component of the MTS sample, is a univariate time series sample that represents a distinct aspect of the observed object. For example, in an intensive care unit (ICU), Patient Monitoring can detect dynamically several physiological parameters, including respiration, ECG, blood pressure, body temperature and the saturation level of blood oxygen over a time interval. Another example, a 2-variate time series sample of ECG (ElectroCardiogram) generated by two leads during an interval is shown in Fig. 1, the data of the first lead is plotted as the solid line while that of the second one is plotted as the dashed line.

MTS classification is an important problem in time series data mining. Because of its multiple variables and the possibility of different lengths for different components, MTS is difficult for traditional machine learning algorithms to address. Recently, a large amount of research has been performed, and many efficient models and techniques have been represented for the classification of multivariate time series [1–3], such as pattern mining [4–6], classification methods [7–11], and similarity measures [12–16]. At the same time, early classification of time series, which means to classify time series data as early as possible provided that the classification quality meets the demand, is an interesting and challenging topic and has attracted a substantial amount of attention [17–19]. For example, an efficient 1-nearest neighbor classification method [18] was proposed to make early prediction on univariate time series, and at the same time, it retained accuracy that was comparable to that of a 1NN classifier using the full-length time series.

However, early classification of multivariate time series is still an open problem that is useful for real data. For example, analyzing the multivariate time series generated by Patient Monitoring and identifying its abnormalities as early as possible could offer doctors an emergency alarm.

To the best of our knowledge, the problem of early classification on MTS data largely remains untouched except for [20]. In [20] Ghalwash et al. defined a multivariate shapelet, which is composed of multiple segments, and each segment is extracted from exactly one component. For this type of shapelet, different starting and ending positions are not allowed in a segment of each component, in other words, all segments of a multivariate shapelet should be extracted in the same sliding time window at the same time. Since interesting patterns often have different intervals for each variable, multivariate shapelets are incapable of including distinctive patterns of all variables unless their lengths are sufficiently long. When a multivariate shapelet is too long, it would not be able to classify the data as early as possible, and the classification costs a substantial amount of storage space and training time. Moreover, this method cannot handle different components with different lengths. For most real-world problems, we cannot expect each component of a MTS object to be equal in length.

Moreover, time series data with the same class label are often composed of various sub-clusters, or sub-concepts, that is to say, samples of a class are collected from different sub-concepts.

* Corresponding author.
E-mail address: glhe@whu.edu.cn (G. He).

http://dx.doi.org/10.1016/j.neucom.2014.07.056
0925-2312/© 2014 Elsevier B.V. All rights reserved.

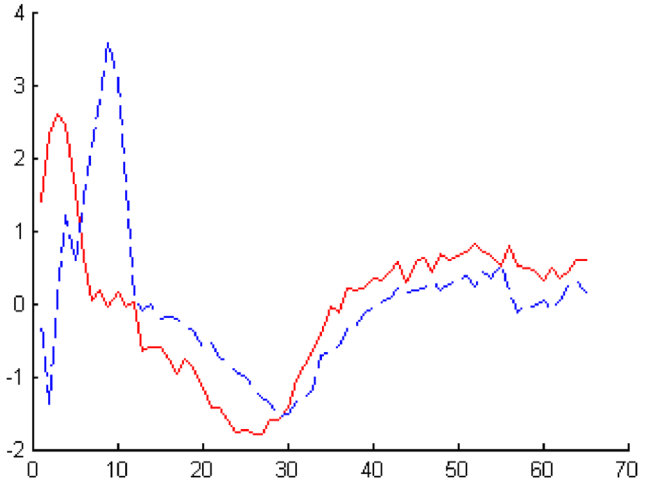

Fig. 1. A sample of ECG MTS generated from 2 sensors.

Since these sub-concepts do not always contain the same number of examples, and this within-class imbalance is implicit in most cases. Therefore, in order to improve the performance of classification, how to deal with the issue of sub-concepts is a challenge.

In this paper, we make concrete progress in answering the above questions as well as discovering the internal relationships among variables. We introduce a novel strategy to evaluate the quality of a shapelet, and develop MCFEC (Mining Core Feature for Early Classification) to obtain core features for each variable independently with a clustering method. Then, two classification methods are introduced for early classification on multivariate time series based on core features. Our contribution can be summarized as follows:

- To discover the internal characteristics of MTS data and enhance the interpretability of classification, we extract feature candidates of each variable independently.
- To deal with within-class imbalance issue a MCFEC method including feature selection and evaluation strategies is proposed to mine core features, which are used to build two classifiers to identify unlabeled MTS data in an explainable way.
- Experimental results clearly show that our proposed method is competitive with the state-of-art methods.

The remainder of this paper is organized as follows. In Section 2, we review the problem definition and related work. Section 3 introduces concrete algorithms to efficiently mine core features. Section 4 discusses two methods of early classification using these core features as a tool. In Section 5, we perform a comprehensive set of experiments on various problems of different domains. Finally, we conclude our work and suggest directions for future work in Section 6.

2. Background and related work

2.1. Background

In this section, we define shapelets and the notations used in this paper.

Definition 1. *Univariate time series*: a univariate time series $s=t_1, t_2,\ldots,t_L$ is an ordered set of L real-valued readings, and L is defined as the length of the time series s (for short, Length(s)=L). For instance, a univariate time series s_0=(1.2, 2.2, 3,6, 1.3, 5.3, 7.1).

Definition 2. *Multivariate time series*: a multivariate time series is a vector of sequences $X=(x_1, x_2,\ldots, x_T)$, where each component x_j is a univariate time series, and the lengths of different components might not be equal.

The MTS object X has T variables, and the corresponding component of the i^{th} variable is x_i.

For the sake of simplicity, we will use the word "time series" as univariate time series, which is a component of a multivariate time series.

Definition 3. *Subsequence*: Given a time series s of length L, s_sub=$s[m, m+n-1]$, is a subsequence of length $n<L$ that has a contiguous position from S starting at the mth position and ending at $(m+n-1)$th position, in other words, s_sub=$t_m,\ldots,t_{m+n-1}$ for $1\le m\le L-n+1$.

Definition 4. *Similarity degree*: For two time series b and s (assuming that $|b|\le|s|$), the similarity degree between b and s is calculated by Sim(b,s)= min{dist(b,s_i)}, where s_i is any subsequence of a time series s with $|s_i|=|b|$, and dist(b,s_i) is Euclidean distance between b and s_i. That is, for two time series samples b and c with equal length L, $\mathrm{dist}(b,c)=\sqrt{\sum_{i=1}^{L}(b_i-c_i)^2}$.

From this definition, we can see that the smaller the value of Sim (b,s) is, the higher the similarity degree between b and s.

Definition 5. *Shapelet or Feature*: A shapelet (feature) is a time series subsequence that is representative of a class. Informally, a shapelet (feature) $p=(b,\delta,c)$, where b is a subsequence, δ is a threshold, and c is a class label. An unknown time series object s is considered to be matching a shapelet p and is labeled as the class of this shapelet if Sim(p,s)$\le\delta$.

For simplicity, later Sim(p, s) is used to represent the similarity degree between the shapelet p and the time series s. Class(s) means the class of a time series s.

Definition 6. *Precision*: Given the time series data D and a feature $p=(b,\delta, c)$, the precision of p is the ratio of the number of samples that have the class label c and could match the feature p and the number of samples that could match the feature p in data D.

$$\mathrm{Precision}(p)=\frac{||\{s|\mathrm{Sim}(s,p)<\delta\ \mathrm{class}(s)=c)\}||}{||\{s|\mathrm{Sim}(s,p)<\delta\}||}, s\in \mathrm{D} \quad (1)$$

Example 1. Given the two-class time series dataset $D=\{s_1(+), s_2(+), s_3(+), s_4(-), s_5(-), s_6(+)\}$ ("+" means a sample belongs to the positive class and "−" presents that it belongs to the negative class), the length of each time series is 12. For a feature $f_1(b_1,0.8,$ positive), we could calculate the similarity degree between the feature f_1 and each sample in D as follows: Sim(f_1,s_1)=0.9 (>0.8), Sim(f_1,s_2)=0.5(<0.8), Sim(f_1,s_3)=1.2(>0.8), Sim(f_1,s_4)=0.6 (<0.8), Sim(f_1,s_5)=1.1(>0.8) and Sim(f_1,s_6)=0.7(<0.8). It is obvious that the number of positive samples that match the feature f_1 (the similarity degree is smaller than the threshold of the feature f_1) is 2, and the number of samples that match the feature f_1 is 3 in the dataset D. Therefore, Precision(f_1)=2/3.

Definition 7. *Recall*: Given time series data D and a feature $p=(b,\delta, c)$, the recall of p is the ratio of the number of samples that have the class label c and could match the feature p and the number of samples that have the class label c in data D.

$$\mathrm{Recall}(p)=\frac{||\{s|\ \mathrm{Sim}(s,p)<\delta\ \mathrm{class}(s)=c)\}||}{||\{s|\ \mathrm{class}(s)=c\}||}, s\in D \quad (2)$$

The recall of p is the true label rate of this feature belonging to class c.

Continuing the above example, the number of positive samples in dataset D is 4; as a result, Recall(f_1)=2/4=0.5.

Download English Version:

https://daneshyari.com/en/article/409780

Download Persian Version:

https://daneshyari.com/article/409780

Daneshyari.com