The Journal of Systems and Software 127 (2017) 179–194

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

Repeating patterns as symbols for long time series representation☆

Jakub Sevcech, Maria Bielikova*

Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Ilkovičova 2, 842 16 Bratislava, Slovakia

ARTICLE INFO

Article history:
Received 1 October 2015
Revised 16 April 2016
Accepted 4 June 2016
Available online 6 June 2016

Keywords:
Time series representation
Symbolic representation
Stream processing
Lower bound

ABSTRACT

Over the past years, many representations for time series were proposed with the main purpose of dimensionality reduction and as a support for various algorithms in the domain of time series data processing. However, most of the transformation algorithms are not directly applicable on streams of data but only on static collections of the data as they are iterative in their nature. In this work we propose a symbolic representation of time series along with a method for transformation of time series data into the proposed representation. As one of the basic requirements for applicable representation is the distance measure which would accurately reflect the true shape of the data, we propose a distance measure operating on the proposed representation and lower bounding the Euclidean distance on the original data. We evaluate properties of the proposed representation and the distance measure on the UCR collection of datasets. As we focus on stream data processing, we evaluate the properties and limitations of the proposed representation on very long time series from the domain of electricity consumption monitoring, simulating the processing of potentially unbound data stream.

1. Introduction

Many different time series representations were proposed over the past years (Esling and Agon, 2012). However, only small portion of them is applicable on stream data processing as most of the transformation procedures are iterative in their nature or they require some sort of statistical information about the whole dataset.

Our primary motivation is to propose a time series representation applicable in stream data processing, in domains where very long (potentially infinite) time series are produced and where repeating shapes are occurring in the course of the time series. The primary application we had in mind when we proposed the representation is forecasting and anomaly detection in data such as counting metrics running on production or consumption data streams, where strong seasonal patterns are occurring. Our prime requirement for such a time series representation is incremental procedure of the data transformation and symbolic representation of reoccurring patterns.

In our work, we are most interested in symbolic representations of equally spaced time series as they enable the application of methods that are not directly applicable on real-valued data (Lin et al., 2007) such as Markov models, suffix trees or many algorithms from the domain of text processing. An example of such representation is SAX (Lin et al., 2007) one of the most widely used time series representations. Similarly to the majority of other representations, however, transformation into the SAX representation is iterative and cannot be directly applicable to stream data processing as it requires statistical information about the whole transformed dataset. Examples of other symbolic time series representation can be found in Lin et al. (2007); Das et al. (1998); Baydogan and Runger (2014); Bagnall et al. (2006), but they all share the same limitation, stream data cannot be directly transformed into these representations.

The representation we propose is based on the symbolic time series representation used in Das et al. (1998) for rule discovery in time series. Clusters of similar subsequences are used as symbols in the transformation of time series into sequences of symbols. This work was influencing many researchers for several years, but they found its two major limitations:

- It is iterative due to the K-means algorithm used for cluster formation.

☆ This paper is based on J. Sevcech and M. Bielikova, Symbolic Time Series Representation for Stream Data Processing published in the proceedings of the 1st IEEE International Workshop on Real Time Data Stream Analytics (part of the BigDataSE-15 conference) (Sevcech and Bielikova, 2015).

* Corresponding author. Fax: +421265420587.

E-mail addresses: jakub.sevcech@stuba.sk (J. Sevcech), maria.bielikova@stuba.sk (M. Bielikova).

http://dx.doi.org/10.1016/j.jss.2016.06.008
0164-1212/© 2016 Elsevier Inc. All rights reserved.

- It has been proved that the transformation process produces meaningless clusters that do not reliably reflect the data they were formed from Keogh and Lin (2004).

In our work, we address both of these limitations. To be able to transform potentially infinite data streams into the proposed representation, we use an incremental greedy clustering algorithm creating new clusters every time new sequence, sufficiently distant from all other clusters, occurs. In previous works multiple authors used various techniques to form meaningful subsequence clusters. Most of these methods limit the number of sequences used in the clustering process by using motifs (Chen, 2007) or perceptually important points (Fu et al., 2005). All of these works used the K-means algorithm in cluster formation. We hypothesize, that not by limiting the number of formed clusters, but by changing the clustering algorithm, we will be able to form meaningful clusters.

According to the authors of another study (Lin et al., 2007) many symbolic time series representations were proposed, but the distance measures on these representations show little correlation with the distance measures on original data. To show our representation is not the case, we propose the distance measure *SymD* that returns the minimum distance between time series in the representation and we show it lower bounds the Euclidean distance on the original time series. To evaluate the applicability of time series representation we use the tightness of lower bounds (TLB) (Keogh et al., 2001) as it is the current consensus in the literature (Wang et al., 2013).

As the majority of existing time series representations focus on processing of static collections of data and we propose our representation to be applicable in stream data processing domain, we evaluate the properties of the proposed representation on static collections of data as well as on very long time series substituting the potentially infinite data streams.

The rest of the paper is organized as follows. Section 2 introduces the symbolic time series representation. Section 3 defines the distance measure on the proposed representation and provides the proof it lower bounds the Euclidean distance on the original data. An experimental evaluation of properties of the proposed representation and distance measure on the number of datasets is presented in Section 4. We conclude by summarizing obtained results and by hints on future work.

2. The symbolic representation

As a base for our time series representation we use an assumption presented in Giannella et al. (2003). The authors state that frequent patterns extracted from time series data are more stable than the time series itself. We use this assumption to form the main idea of our representation as to represent time series data as a sequence of reoccurring patterns. We search for reoccurring similar subsequences in the course of the whole data stream by clustering subsequences. We transform them into sequences of symbols where every subsequence cluster identifier is transformed into a symbol similarly to the representation proposed in Das et al. (1998). For the purpose of our work, we will refer the proposed representation as to *Incremental Subsequence Clustering* (*ISC*).

The transformation of stream data to the *ISC* representation can be divided into three steps:

1. Split incoming data into overlapping subsequences using running window.
2. Cluster z-normalized subsequences by their similarity.
3. Use cluster identifiers as symbols, subsequences are transformed to. In connection with normalization coefficients, these symbols approximate the original data.

As the processed time series may contain some levels of noise and trend, the preprocessing step may be introduced into the transformation. To remove the noise present in the formed subsequences and to highlight important parts of the data, some level of smoothing can be applied before the symbol formation as the introduction of smoothing before the symbols are created can produce more stable alphabet of symbols. To find the correct level of smoothing, one could use a framework such as the one presented in Miao et al. (2015), based on Minimum Description Length principle (Grünwald, 2007). In the evaluation of the proposed representation presented in this paper however, we did not use any smoothing as we did not want to introduce any error by omitting minor changes in the shape of the processed time series.

The ISC representation is inspired by the representation presented in Das et al. (1998), with two important differences:

- we use overlapping symbols and
- we don't use K-means algorithm in symbol formation.

The redundancy contained in overlapping symbols could be used to improve the reconstruction accuracy when transforming data back to their raw form, and to some extent it is used in the similarity measure on the data transformed into ISC (presented in later parts of the paper). The main motivation to introduce the overlapping symbols however, is to support one of intended applications of the time series representation - short term time series forecasting. If time series is transformed into a symbolic representation with overlapping symbols incrementally, in every moment at least the length of the overlapping part of two symbols could be used to search for similar shapes in alphabet of symbols. The last part of the processed time series could be simply compared to early parts of symbols in the alphabet. The later part of the most similar symbol from the alphabet can be then used to forecast the rest of the symbol's length. Of course, this would be just the simplest method which could be extended by employing other similar symbols or sequence of symbols occurring earlier in the transformed time series.

The main difference of the proposed *ISC* representation to the representation Das et al. (1998) used is the clustering algorithm we use for symbol formation. They used K-means, which is iterative in its nature and requires the number of formed clusters to be specified in advance. As shown in Keogh and Lin (2004), this results in meaningless cluster formation as the cluster centre does not reflect the data, cluster is formed from, but transforms into a shifted sinusoidal shape regardless the shape of the transformed data. We chose different approach to symbol formation by not using K-menas clustering algorithm.

We use incremental greedy algorithm not limiting the number of clusters but limiting the maximal distance of subsequences from the cluster centre. The algorithm assigns subsequence into the cluster if its distance from the cluster centre is smaller than the predefined threshold (referenced as limit distance). The algorithm forms new cluster with the subsequence in its centre if no cluster with the distance to the processed subsequence lower than the maximal distance exists.