Pattern Recognition 42 (2009) 3015 – 3031

Contents lists available at ScienceDirect

Pattern Recognition

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

On-line motif detection in time series with SwiftMotif

Erich Fuchs[a], Thiemo Gruber[b], Jiri Nitschke[b], Bernhard Sick[b,*]

[a]Institute for Software Systems in Technical Applications (Forwiss), University of Passau, Germany
[b]Computationally Intelligent Systems Group, Department of Informatics and Mathematics, University of Passau, Germany

ARTICLE INFO

Article history:
Received 21 August 2008
Received in revised form 7 February 2009
Accepted 4 May 2009

Keywords:
Temporal data mining
Time series
Motif detection
Polynomial approximation
Orthogonal polynomials
Probabilistic modeling
Piecewise polynomial representation
Piecewise probabilistic representation
Segmentation
SwiftMotif

ABSTRACT

This article presents SwiftMotif, a novel technique for on-line motif detection in time series. With this technique, frequently occurring temporal patterns or anomalies can be discovered, for instance. The motif detection is based on a fusion of methods from two worlds: probabilistic modeling and similarity measurement techniques are combined with extremely fast polynomial least-squares approximation techniques. A time series is segmented with a data stream segmentation method, the segments are modeled by means of normal distributions with time-dependent means and constant variances, and these models are compared using a divergence measure for probability densities. Then, using suitable clustering algorithms based on these similarity measures, motifs may be defined. The fast time series segmentation and modeling techniques then allow for an on-line detection of previously defined motifs in new time series with very low run-times. SwiftMotif is suitable for real-time applications, accounts for the uncertainty associated with the occurrence of certain motifs, e.g., due to noise, and considers local variability (i.e., uniform scaling) in the time domain. This article focuses on the mathematical foundations and the demonstration of properties of SwiftMotif—in particular accuracy and run-time—using some artificial and real benchmark time series.

1. Introduction

Motifs are approximately repeated subsequences of time series [1]. Typical tasks where such motifs must be detected are (cf. [1–6]), for example,

- mining of temporal association rules,
- novelty or anomaly detection within time series,
- summarization and indexing of time series,
- time series forecasting, or
- clustering and classification of time series.

Applications for motif detection algorithms can be found in many fields, e.g., medicine, bioinformatics, meteorology, music analysis, or robotics. The problem of motif detection has been addressed since several years also using related terms such as temporal pattern, episode, subsequence, or shape instead of motif.

In this article we introduce *SwiftMotif*, a new technique for the on-line detection of motifs in continuous time series. SwiftMotif first supports the definition of motifs, e.g., frequently occurring temporal patterns or anomalies, and then finds these motifs in new time series even if they are distorted by noise or scaled in the time domain. SwiftMotif is particularly suited for time series that can be segmented at distinct points (cf. the concept of perceptually important point in Ref. [7]). In this article we lay the mathematical foundations of the technique and investigate its properties with a few application examples. SwiftMotif offers the following key advantages which are needed in many applications scenarios (cf., e.g., [1–3,8]):

- SwiftMotif is a *data stream algorithm* [9]. Therefore, it is suitable for real-time applications and also for applications with very large datasets.
- SwiftMotif models and considers the *uncertainty of the occurrence of motifs*. Thus, it is able to deal with the detection of motifs in the presence of noise.
- SwiftMotif allows *local variability* of the time series. That is, it considers linear (uniform) scaling of the motifs in the time domain.

The key concept that leads to these desirable properties is a fusion of probabilistic modeling techniques with extremely fast polynomial least-squares (LS) approximation techniques. This leads to a new kind of motif representation and a new kind of (dis-)similarity (distance) measure for motifs. The former is basically a normal distribution with time-dependent, polynomial-shaped mean and constant variance. The latter is based on a divergence measure for probability densities.

* Corresponding author.
E-mail addresses: fuchse@forwiss.uni-passau.de (E. Fuchs), grubert@fim.uni-passau.de (T. Gruber), nitschke@fim.uni-passau.de (J. Nitschke), sick@fim.uni-passau.de (B. Sick).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.05.004

The remainder of this article is structured as follows: Section 2 presents some related work in the field of motif detection. In Section 3 we outline the idea of SwiftMotif in some more detail to motivate the various mathematical foundations introduced in the following sections. Section 4 presents the foundations in the field of probability theory and statistics, Section 5 sets out the foundations in the field of least-squares approximation with certain bases of orthogonal functions, and Section 6 fuses the two worlds to define an online segmentation algorithm which is the basis for motif detection. Then, Section 7 introduces SwiftMotif, which is based on the representation and dissimilarity measures introduced before. Section 8 presents some experimental results on artificial as well as on real-world data. The properties (in particular accuracy and run-time) of our approach are evaluated in detail. Finally, Section 9 summarizes the major findings and gives an outlook to possible future work.

2. Related work in the field of motif detection

Techniques for finding and extracting new motifs from given data and detecting previously known motifs are used in many applications (cf. [2]), for example, to identify and represent events in the field of sensor data analysis [10,11], to find unknown patterns in financial datasets [12], or to classify chemical reactions or genes in the field of chemical and biomedical engineering [13,14]. An overview of various data mining applications where motif extraction is utilized can be found in Refs. [15,16], for example. Often, the extraction of new motifs is done in an unsupervised manner based on finding frequently occurring patterns by segmenting and clustering the data (cf. [17]).

Obviously, one of the key issues in motif detection is the representation of the time series subsequences. Various approaches, such as the minimum description length principle [18] or HMM [19,20] have been used to represent motifs in time series data. Also piecewise linear representation, cf. [21–23], or piecewise polynomial representation [24–26] is frequently used to describe the characteristics of time series subsequences.

Another key issue in motif detection is an appropriate distance or (dis-)similarity measure used for clustering and comparing new data with previously determined motifs. Based on the representation of the time series subsequences, many different measures such as the Euclidean distance, aligned subsequences, or autocorrelation functions have been used (cf. [7,27,28], for instance).

For on-line applications such as motion analysis, motifs are often used to extract and represent significant parts of the data [29]. Araki et al. [30] successfully improved human motion analysis by extracting frequent patterns from motion information, Tanaka et al. [3] computed motifs from multivariate time series representing the three-dimensional movement of body parts in order to get information about repeated motion sequences of humans, and Celly and Zordan [8] used motifs to identify human movements and applied it to the animation of people textures and a so-called real-time human proxy.

In this article, we are addressing this problem of on-line motif detection independent from a particular application scenario. The goal is to provide a technique that has the same temporal complexity as techniques that are based on relatively simple subsequence representations (e.g., piecewise linear models) and similarity measures (e.g., Euclidean distance), but a higher accuracy. We can state that a technique based on representation forms and dissimilarity measures similar to those proposed in this article has not been investigated yet.

3. Outline of the solution

In this section we will sketch the basic ideas and the operation principle of SwiftMotif.

At first, we assume that a univariate, continuous time series can be regarded as a mixture of deterministic and random (probabilistic) components. Our model of the time series is a distribution

$$p(y|x, \mathbf{w}, \rho), \tag{1}$$

with variables $x \in \mathbb{R}$, which models the time, and $y \in \mathbb{R}^+$. The *weight vector* $\mathbf{w} \in \mathbb{R}^{K+1}$ (with $K \in \mathbb{N}_0$) is a coefficient (parameter) vector of the deterministic part of the model, whereas the *precision* $\rho \in \mathbb{R}^+$ is a parameter of the probabilistic part.

We assume further that the deterministic component can be described by means of a function of a certain type, in our case a polynomial. As the data we observe are assumed to be noisy, we need the probabilistic component. Here, we use a normal distribution to model white noise.

Thus, our model of a time series can be regarded as a normal distribution, where the time-dependent mean is given by the deterministic component:

$$p(y|x, \mathbf{w}, \rho) = \mathcal{N}(y|f_{\mathbf{w}}(x), \rho^{-1}). \tag{2}$$

The precision ρ is the inverse of the variance of this normal distribution. The variable x models the time and the mean $f_{\mathbf{w}}$ is a polynomial of degree K given by a linear combination (with weight vector $\mathbf{w}$) of certain basis polynomials. Fig. 1 shows an example for a time series modeled by polynomials of degree 0, 1, and 2, respectively. The shaded area is determined by the variance which is equivalent to the (average) least-squares error of the polynomial approximation. It can be stated that the parabola in Fig. 1(c) fits this time series quite well.

We aim at detecting motifs, i.e., recurring temporal patterns, with very low computational effort. Therefore, we will introduce techniques that allow to determine $\mathbf{w}$ and ρ extremely fast in *growing* or *sliding time windows* (referred to as GW and SW techniques, respectively). Assume that we are already given a probabilistic model for a time series as shown in Fig. 1(c). Then, a new sample is observed, see Fig. 2(a). We will show how a new model in the GW case—Fig. 2(b)—and a new model in the SW case—Fig. 2(c)—can be determined from the solution shown in Fig. 2(a) with computational costs being independent from the length of the time window (i.e., the number of samples observed). For the example in Fig. 2 it can be stated that in both cases—GW and SW—the precision gets lower when the new observation is considered.

Next, we assume that motifs within a time series begin and end at characteristic points such as maxima, minima, or inflection points. To find these points, we use a time series segmentation technique which is based on information produced by the GW and/or SW modeling techniques. For example, a segmentation criterion may be based on the likelihood of a given probabilistic model for a new observation. If we decide upon the existence of a segmentation point (i.e., a segment boundary) at a certain point in time, a new segment begins, which we immediately start to model using the GW technique. The SW technique is needed in addition to the GW technique to define segmentation criteria. The approach finally leads to a piecewise probabilistic model of a time series as sketched in Fig. 3. This piecewise probabilistic model contains a piecewise polynomial model as well (cf., e.g., [24–26]). Depending on the degree of the polynomials used for modeling, the segments could also be more complex.

A probabilistic model of a segment of the time series (i.e., a subsequence) can now be regarded as an abstract description of a motif candidate. The *shape* of the motif candidate is described by a polynomial of a certain degree (i.e., the time-dependent mean of the probabilistic model) and the *uncertainty* (see, e.g., Ref. [31]) associated with the occurrence of that shape due to noise is described by the variance of the model.

Motifs can now be defined applying suitable clustering algorithms or the like. Therefore, we have to measure the (dis-)similarity of two