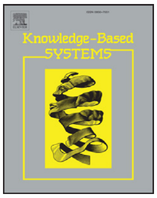
Knowledge-Based Systems 92 (2016) 127–137

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Particle swarm optimization for time series motif discovery

Joan Serrà [a,b,*], Josep Lluis Arcos [b]

[a] *Telefónica Research, Pl. Ernest Lluch i Martín 5, 08019 Barcelona, Spain*
[b] *IIIA-CSIC, Campus de la UAB s/n, 08193 Bellaterra, Spain*

ARTICLE INFO

Article history:
Received 8 July 2015
Revised 15 October 2015
Accepted 15 October 2015
Available online 26 October 2015

Keywords:
Motifs
Time series
Anytime algorithms
Particle swarm optimization
Multimodal optimization

ABSTRACT

Efficiently finding similar segments or motifs in time series data is a fundamental task that, due to the ubiquity of these data, is present in a wide range of domains and situations. Because of this, countless solutions have been devised but, to date, none of them seems to be fully satisfactory and flexible. In this article, we propose an innovative standpoint and present a solution coming from it: an anytime multimodal optimization algorithm for time series motif discovery based on particle swarms. By considering data from a variety of domains, we show that this solution is extremely competitive when compared to the state-of-the-art, obtaining comparable motifs in considerably less time using minimal memory. In addition, we show that it is robust to different implementation choices and see that it offers an unprecedented degree of flexibility with regard to the task. All these qualities make the presented solution stand out as one of the most prominent candidates for motif discovery in long time series streams. Besides, we believe the proposed standpoint can be exploited in further time series analysis and mining tasks, widening the scope of research and potentially yielding novel effective solutions.

1. Introduction

Time series are sequences of real numbers measured at successive, usually regular time intervals. Data in the form of time series pervade science, business, and society. Examples range from economics to medicine, from biology to physics, and from social to computer sciences. Repetitions or recurrences of similar phenomena are a fundamental characteristic of non-random natural and artificial systems and, as a measurement of the activity of such systems, time series often include pairs of segments of strikingly high similarity. These segment pairs are commonly called motifs [34], and their existence is unlikely to be due to chance alone. In fact, they usually carry important information about the underlying system [41]. Thus, motif discovery is fundamental for understanding, characterizing, modeling, and predicting the system behind the time series. Besides, motif discovery is a core part of several higher-level algorithms dealing with time series, in particular classification, clustering, summarization, compression, and rule-discovery algorithms [see, e.g., 40, 41].

Identifying similar segment pairs or motifs typically implies examining all pairwise comparisons between all possible segments in a time series. This, specially when dealing with long time series streams, results in prohibitive time and space complexities. It is for this reason that the majority of motif discovery algorithms resort to some kind of data discretization or approximation that allows them to hash and retrieve segments efficiently. Following the works by Lin et al. [34] and Chiu et al. [13], many of such approaches employ the SAX representation [35] and/or a sparse collision matrix [9]. These allow them to achieve a theoretically low computational complexity, but sometimes at the expense of very high constant factors. In addition, approximate algorithms usually suffer from a number of data-dependent parameters that, in most situations, are not intuitive to set (e.g., time/amplitude resolutions, dissimilarity radius, segment length, minimum segment frequency, etc.).

A few recent approaches overcome some of these limitations. For instance, Castro and Azevedo [11] propose an amplitude multiresolution approach to detect frequent segments, Li and Lin [33] use a grammar inference algorithm for exploring motifs with lengths above a certain threshold, Wilson et al. [55] use concepts from immune memory to deal with different lengths, and Floratou et al. [18] combine suffix trees with segment models to find motifs of any length. Nevertheless, in general, these approaches still suffer from other data-dependent parameters whose correct tuning can require considerable time. In addition, approximate algorithms are restricted to a specific dissimilarity measure between segments (the one implicit in their discretization step) and do not allow easy access to preliminary results, which is commonly known as anytime algorithms

[*] Corresponding author at: Telefónica Research, Pl. Ernest Lluch i Martin 5, 08019 Barcelona, Spain. Tel.: +34 931233010.
E-mail addresses: joan.serra@telefonica.com (J. Serrà), arcos@iiia.csic.es (J.L. Arcos).

http://dx.doi.org/10.1016/j.knosys.2015.10.021
0950-7051/© 2015 Elsevier B.V. All rights reserved.

[58]. Finally, to the best of our knowledge, only the authors in [51,52,56] consider the identification of motif pairs containing segments of different lengths. This can be considered a relevant feature, as it produces better results in a number of different domains [56].

In contrast to approximate approaches, algorithms that do not discretize the data have been comparatively much less popular, with low efficiency generally. Exceptions to this statement achieved efficiency by sampling the data stream [12] or by identifying extreme points that constrained the search [38]. In fact, until the work of Mueen and Keogh [44], the exact identification of time series motifs was thought to be intractable for even time series of moderate length. In said work, a clever segment ordering was combined with a lower bound based on the triangular inequality to yield the true, exact, most similar motif. According to the authors, the proposed algorithm was more efficient than existing approaches, including all exact and many approximate ones [44]. After Mueen et al.'s work, a number of improvements have been proposed, the majority focusing on eliminating the need to set a fixed segment length [39,45,57].

Mueen himself has recently published a variable-length motif discovery algorithm which clearly outperforms the iterative search for the optimal length using the algorithm in [44] and, from the reported numbers, also outperforms further approaches as in [39,45,57]. This algorithm, called MOEN [40], is essentially parameter-free, and is believed to be one of the most efficient motif discovery algorithms available nowadays. However, its execution time may still be unaffordable in a number of situations. Furthermore, MOEN is specifically designed to work with Euclidean distances after z-normalization. In general, exact motif discovery algorithms have important restrictions with regard to the dissimilarity measure, and many of them still suffer from being non-intuitive and tedious to tune parameters. Moreover, few of them allow for anytime versions and, to the best of our knowledge, not one of them is able to identify motif pairs containing segments of different lengths. With the approach we propose here we try to overcome all these shortcomings at the same time.

In this article, we propose a new standpoint to time series motif discovery by treating the problem as an anytime multimodal optimization task. To the best of our knowledge, this standpoint is completely unseen in the literature. We first motivate such a standpoint and discuss its multiple advantages (Section 2). Next, we present SWARMMOTIF (Section 3), an anytime algorithm for time series motif discovery based on particle swarm optimization (PSO). We subsequently evaluate the performance of the proposed approach using 9 different real-world time series from distinct domains (Section 4). These include economics, car traffic, entomology, medical data, audio, climate, and power consumption. Our results show that SWARMMOTIF is extremely competitive when compared to the state-of-the-art, obtaining motif pairs of comparable similarity in considerably less time and with minimum storage requirements (Section 5). Moreover, we show that SWARMMOTIF is significantly robust against different implementation choices. These two aspects, together with its flexibility and extension capabilities, make SWARMMOTIF a unique novel solution for time series motif discovery. The latter implies that SWARMMOTIF can, for instance, deal with motifs of different lengths, apply uniform scaling, use any suitable dissimilarity measure, or incorporate notions of motif frequency. To conclude, we briefly comment on the application of multimodal optimization techniques to time series analysis and mining, which we believe has great potential (Section 6). The data and code used in our experiments are available online[1].

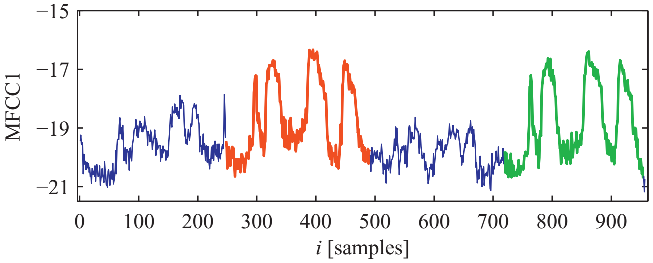


Fig. 1. Example of a time series motif pair found in the WILDLIFE time series of [42] using SWARMMOTIF and normalized dynamic time warping as the dissimilarity measure: $a = 248$, $w_a = 244$, $b = 720$, and $w_b = 235$.

2. Time series motif discovery as an anytime multimodal optimization task

2.1. Definitions and task complexity

From the work by Mueen et al. [40,44], we can derive a formal, generic similarity-based definition [41] of time series motifs. Given a time series $\mathbf{z}$ of length n, $\mathbf{z} = [z_1, \dots z_n]$, a normalized segment dissimilarity measure D, and a temporal window of interest between $w_{\min}$ and $w_{\max}$ samples, the top-k time series motifs $\mathcal{M} = \{\mathbf{m}_1, \dots \mathbf{m}_k\}$ correspond to the k most similar segment pairs $\mathbf{z}_a^{w_a} = [z_a, \dots z_{a+w_a-1}]$ and $\mathbf{z}_b^{w_b} = [z_b, \dots z_{b+w_b-1}]$, for w_a, w_b $\in [w_{\min}, w_{\max}]$, $a \in [1, n - w_a + 1]$, and $b \in [1, n - w_b + 1]$. Thus, we see that the ith motif can be fully described by the tuple $\mathbf{m}_i = \{a, w_a, b, w_b\}$. To avoid so-called trivial matches [34], we can force that motifs are non-overlapping[2], that is, $a + w_a < b$ or $b + w_b < a$. The motifs in $\mathcal{M}$ are ordered from lowest to highest dissimilarity such that $D(\mathbf{m}_1) \le D(\mathbf{m}_2) \le \dots \le D(\mathbf{m}_k)$ where $D(\mathbf{m}_i) = D(\{a, w_a, b, w_b\}) = D(\mathbf{z}_a^{w_a}, \mathbf{z}_b^{w_b})$. An example of a time series motif pair from a real data set is shown in Fig. 1.

It is important to stress that D needs to normalize with respect to the lengths of the considered segments. Otherwise, we would not be able to compare motifs of different lengths. There are many ways to normalize with respect to the length of the considered segments. Ratanamahatana and Keogh [49] list a number of intuitive normalization mechanisms for dynamic time warping that can easily be applied to other measures. For instance, in the case of a dissimilarity measure based on the L_p norm, we can directly divide by the segment length[3], using brute-force upsampling to the largest length when $w_a \neq w_b$.

From the definitions above, we can see that a brute-force search in the motif space for the most similar motifs is of $O(n^2 {w_\Delta}^2)$, where $w_\Delta = w_{\max} - w_{\min} + 1$ (for the final time complexity one needs to further multiply by the cost of calculating D). Hence, for instance, in a perfectly feasible case where $n = 10^7$ and $w_\Delta = 10^3$, we have 10^{20} possibilities. Magnitudes like this challenge the memory and speed of any optimization algorithm, specially if we have no clue to guide the search [23]. However, it is one of our main objectives to show here that time series generally provide some continuity to this search space, and that this continuity can be exploited by optimization algorithms.

2.2. Continuity

A fundamental property of time series is autocorrelation, implying that consecutive samples in a time series have some degree of

[1] http://www.iiia.csic.es/~jserra/swarmmotif

[2] Notice that, following [40], this definition can be trivially extended to different degrees of overlap.

[3] The only exception is with L_∞, which could be considered as already being normalized.