Contents lists available at ScienceDirect

Information Sciences

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

# Predefined pattern detection in large time series

Shengfa Miao[a,b,*], Ugo Vespier[b], Ricardo Cachucho[b], Marvin Meeng[b], Arno Knobbe[b]

[a] Lanzhou University, Lanzhou, China
[b] Liacs, Leiden University, The Netherlands

ARTICLE INFO

*Article history:*
Received 1 April 2014
Received in revised form 5 April 2015
Accepted 11 April 2015
Available online xxxx

*Keywords:*
Pattern detection
Landmark
Trust region
MDL
Predefined pattern
Time series representation

ABSTRACT

Predefined pattern detection from time series is an interesting and challenging task. In order to reduce its computational cost and increase effectiveness, a number of time series representation methods and similarity measures have been proposed. Most of the existing methods focus on full sequence matching, that is, sequences with clearly defined beginnings and endings, where all data points contribute to the match. These methods, however, do not account for temporal and magnitude deformations in the data and result to be ineffective on several real-world scenarios where noise and external phenomena introduce diversity in the class of patterns to be matched. In this paper, we present a novel pattern detection method, which is based on the notions of templates, landmarks, constraints and trust regions. We employ the Minimum Description Length (MDL) principle for time series preprocessing step, which helps to preserve all the prominent features and prevents the template from overfitting. Templates are provided by common users or domain experts, and represent interesting patterns we want to detect from time series. Instead of utilising templates to match all the potential subsequences in the time series, we translate the time series and templates into landmark sequences, and detect patterns from landmark sequence of the time series. Through defining constraints within the template landmark sequence, we effectively extract all the landmark subsequences from the time series landmark sequence, and obtain a number of landmark segments (time series subsequences or instances). We model each landmark segment through scaling the template in both temporal and magnitude dimensions. To suppress the influence of noise, we introduce the concept of *trust region*, which not only helps to achieve an improved instance model, but also helps to catch the accurate boundaries of instances of the given template. Based on the similarities derived from instance models, we introduce the probability density function to calculate a similarity threshold. The threshold can be used to judge if a landmark segment is a true instance of the given template or not. To evaluate the effectiveness and efficiency of the proposed method, we apply it to two real-world datasets. The results show that our method is capable of detecting patterns of temporal and magnitude deformations with competitive performance.

* Corresponding author at: Liacs, Leiden University, The Netherlands.

*E-mail addresses:* miaoshengfa@hotmail.com (S. Miao), u.vespier@liacs.leidenuniv.nl (U. Vespier), r.cachucho@liacs.leidenuniv.nl (R. Cachucho), m.meeng@liacs.leidenuniv.nl (M. Meeng), a.j.knobbe@liacs.leidenuniv.nl (A. Knobbe).

http://dx.doi.org/10.1016/j.ins.2015.04.018
0020-0255/© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction

This paper focuses on the problem of detecting instances of predefined patterns from time series [15,56]. While most pattern detection algorithms in time series deal with discovering previously unknown, frequently recurring regularities in the data, here we assume that one or more example sequences (the *templates*) are provided by a domain expert, and instances of these need to be identified in the actual data. During this detection, one needs to allow for a certain degree of difference between the template and the instances, for example because the instance is somewhat longer or shorter in duration, the magnitude of the signal is different, or parts of the signal are either stretched or compressed in time (so-called warps).

Li Wei et al. [56] mention a number of use-cases that motivate the predefined pattern detection problem. For example, in ECG monitoring, a cardiologist may observe some interesting pattern that he or she wants to annotate, and flag any future occurrences, to be investigated by the cardiologist or fellow experts. Alternatively, in insect pest control, one would like to observe specific cases of harmful insects, as identified by specific patterns of audio signal (wing beats). In our application to infrastructure monitoring, the predefined pattern detection problem is relevant for specifying and detecting known disturbances in the data, that can then be removed from the signal, or accounted for in subsequent modelling steps. For example, when monitoring the structural health of a bridge, the measured signal is dominated by recurring and understandable peaks due to vehicles crossing the bridge and traffic jams. One can imagine an expert providing a template for each of these phenomena, after which all instances should be identified, regardless of the speed and weight of the vehicles (influencing the width and height of the hump in the signal), or the duration of the traffic jam.

When matching a predefined phenomenon (a template [40,41,48]) with the time series under investigation, it is not always required to involve every individual measurement in the selected interval and in the template. In fact, when a certain level of fuzzy matching is required, it makes sense to somehow simplify the signal, or extract some key features that are characteristic for the sequence in question. This condensed representation can then be used to compare the time series with the template, both effectively (the matching is only based on the characteristic aspects) and efficiently (no computation is wasted on insignificant details). Specifically when large time series with high sampling rates are concerned, and the matching is nontrivial due to warps, efficient representation methods can be helpful. A considerable number of such methods have been proposed in the past, including Symbolic Aggregate approXimation (SAX) [29], bit-level approximation [7], and Piecewise Aggregate Approximation (PAA) [25].[1] In this paper specifically, we focus on the representation of time series by means of *landmarks* [43] (also referred to as key-points [8], break-points [47] and change-points [37]), which can be thought of as those points in the time series that are obviously remarkable (peaks, valleys, inflection points, …). Rather than matching every detail of the data and the template, only the landmarks will be matched, and subsequent landmarks will be checked for their relationship to one another.

We match the given template to the actual data in three steps. The first step involves transforming the time series into a sequence of landmarks, which preserves all the prominent features. The second step is landmark subsequence selection, which is based on constraints over the landmarks occurring in the templates. The third step is instance model construction, which introduces a *trust region* to model the time series segments corresponding to the selected landmark subsequence. Unlike most of the representation and similarity methods, which are designed mainly for full sequence matching [15], our proposed approach is capable of processing both full sequence and subsequence matching of various length, while being less sensitive to noise, and being able to handle deformations in both magnitude and temporal dimensions.

One of the challenges when extracting landmarks from actual data is the noise and high-frequency vibrations that are included. An obvious step to get rid of such distractions and to produce a set of meaningful landmarks is to convolve the signal with a smoothing kernel. The question now becomes what level of smoothing is appropriate for the template in question. Too much smoothing may cause one to miss characteristic landmarks in the data, and too little smoothing will cause an abundance of landmarks at every little disturbance in the data. We propose an MDL-based solution to this challenge, that picks the correct smoothing level. Minimum Description Length (MDL) [20,49,51] is an information-theoretic model-selection framework that selects the best model according to its ability to *compress* the given data.

The contributions of this paper are summarised as follows:

- It provides a general definition of a template for time series.
- It proposes the use of landmarks: a triple involving temporal, magnitude and type information.
- It takes the relationship between landmarks within a landmark sequence as constraints for landmark subsequence selection.
- It introduces the concept of a trust region from the image processing domain [32] to time series to build a reliable instance model.
- It employs MDL [20,49,51] for selection of the right smoothing level for landmark extraction.

The rest of this paper is organised as follows. Section 2 gives the definitions of template and landmark, and specifies the task of predefined pattern detection. Section 3 introduces the concept of landmark constraints. The question of choosing the right smoothing level through MDL is discussed in Section 4. In Section 5, instance models are used to fit the template to

[1] A comprehensive list of representation methods for time series is given in Section 7.

Please cite this article in press as: S. Miao et al., Predefined pattern detection in large time series, Inform. Sci. (2015), http://dx.doi.org/10.1016/j.ins.2015.04.018