Expert Systems with Applications 42 (2015) 3150–3158

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Local-shapelets for fast classification of spectrographic measurements

Daniel Gordon [a,c,*], Danny Hendler [a], Aryeh Kontorovich [a], Lior Rokach [b,c]

[a] Department of Computer Science, Ben-Gurion University of The Negev, Be'er Sheva 84105, Israel
[b] Department of Information Systems Engineering, Ben-Gurion University of The Negev, Be'er Sheva 84105, Israel
[c] Telekom Innovation Laboratories, Ben-Gurion University of The Negev, Be'er Sheva 84105, Israel

ARTICLE INFO

*Article history:*
Available online 10 December 2014

*Keywords:*
Spectrography
Time series
Classification
Shapelets
Local

ABSTRACT

Spectroscopy is widely used in the food industry as a time-efficient alternative to chemical testing. Lightning-monitoring systems also employ spectroscopic measurements. The latter application is important as it can help predict the occurrence of severe storms, such as tornadoes.

The *shapelet* based classification method is particularly well-suited for spectroscopic data sets. This technique for classifying time series extracts patterns unique to each class. A significant downside of this approach is the time required to build the classification tree. In addition, for high throughput applications the classification time of long time series is inhibitive. Although some progress has been made in terms of reducing the time complexity of building shapelet based models, the problem of reducing classification time has remained an open challenge.

We address this challenge by introducing *local-shapelets*. This variant of the shapelet method restricts the search for a match between shapelets and time series to the vicinity of the location from which each shapelet was extracted. This significantly reduces the time required to examine each shapelet during both the learning and classification phases. Classification based on local-shapelets is well-suited for spectroscopic data sets as these are typically very tightly aligned. Our experimental results on such data sets demonstrate that the new approach reduces learning and classification time by two orders of magnitude while retaining the accuracy of regular (non-local) shapelets-based classification. In addition, we provide some theoretical justification for local-shapelets.



## 1. Introduction

Spectroscopy is a field devoted to the study and characterization of physical systems by measuring the electromagnetic frequencies they absorb or emit (Herrmann & Onkelinx, 1986). Items differing in their chemical composition or molecular bonds absorb or emit light at different wavelengths leaving a different spectroscopic fingerprint thus enabling differentiation between them. For example, Al-Jowder, Kemsley, and Wilson (2002) used mid-infrared spectroscopy to detect meat adulteration by comparing the spectra of adulterated meat with that of unadulterated meat. A study by Briandet, Kemsley, and Wilson (1996) discriminated between two different types of coffee beans (Arabica and Robusta) using mid-infrared spectroscopy. Other methods for distinguishing between different types of food exist, which are based on wet chemical analysis (Bicchi, Binello, Legovich, Pellegrino, & Vanni, 1993; Lumley, 1996; Sharma, Srivastava, Gill, & Joshi, 1994). The advantages of spectroscopy over wet chemical analysis are in its simplicity (Briandet et al., 1996) and speed of response.

Spectroscopic measurements are also generated by systems monitoring lightning (Eads et al., 2002). This application is important as relative percentages of different types of lightning can indicate the outbreak of severe storms, such as tornadoes. In addition to laboratory research, spectroscopic equipment is starting to be mass produced for every day use allowing anyone to analyze their surroundings with the aid of spectroscopic measurements (SCIO, 2014). The measurements are uploaded to a cloud service where they are analyzed and then the results of the analysis are made available. The service is cloud based, requiring algorithms with high throughput to enable a quick response to high volumes of queries by users.

The outcome of the spectroscopic analysis of a physical system is a vector in which each index represents a frequency and each value is the measured intensity of that frequency. The representation of spectroscopic measures and time series are identical (Ye & Keogh, 2011a), as the only explicit data are the measurements

* Corresponding author at: Department of Computer Science, Ben-Gurion University of The Negev, Be'er Sheva 84105, Israel. Tel.: +972 (0)86428782; fax: +972 (0)86477650.

*E-mail addresses:* gordonda@cs.bgu.ac.il (D. Gordon), hendlerd@cs.bgu.ac.il (D. Hendler), karyeh@cs.bgu.ac.il (A. Kontorovich), liorrk@bgu.ac.il (L. Rokach).

http://dx.doi.org/10.1016/j.eswa.2014.11.043
0957-4174/

and the meaning of each measurement is defined by its location in the vector. This equivalence allows the application of time series classification methods to the field of spectroscopy. A previous experimental study (Hills, Lines, Baranauskas, Mapp, & Bagnall, 2013) showed that the shapelet based classification method is particularly suited for data sets from the field of spectroscopy, as it achieved a higher accuracy than other machine-learning classification methods.

Recently, Ye and Keogh (2011a) introduced the shapelets approach for classifying time series. A *shapelet* is a subsequence extracted from one of the time series in the data set. The shapelet is chosen by its ability to distinguish between time series from different classes. A test time series is classified based on its distance from the shapelet. In the case of multiple shapelets, these form the nodes of a classification tree. The intuition behind this approach is that the pattern best separating the classes is not necessarily an entire time series. Rather, a certain subsequence may best describe the discriminating pattern. Ye and Keogh's algorithm considers all possible subsequences in the training set in order to identify those shapelets that yield the optimal split. Through the rest of this paper, we will refer to this algorithm as the *YK-algorithm*.

Two key advantages of classification with shapelets are the accuracy and interpretability of the induced classification model, as it supplies information on the patterns characteristic of the different classes (Ye & Keogh, 2011a). A significant downside of this approach is the time required for building the classification tree (Hills et al., 2013; Mueen, Keogh, & Young, 2011; Rakthanmanon & Keogh, 2013). The search for the best shapelet requires examining *all* subsequences of *all* lengths from *all* the time series in the training set, and for each shapelet calculating its distance to each time series at *all* possible locations. Even for small data sets, this process has a time scale of days, and for large data sets, the time scale becomes one of years. Hence, the original implementation on commonly available hardware is only practical on the smallest of data sets. Additionally, for high-throughput applications, the classification time may be prohibitively expensive for long time series. This is because at each node of the tree, all possible matches between the node's shapelet and the time series to be classified need to be examined.

### 1.1. Our contributions

Our goal was to reduce both learning and classification time without impairing the accuracy of the induced model by exploiting a feature common to spectroscopic data – the localized nature of information in the time series. In the YK-algorithm, no importance is attributed to the location from which the shapelet was extracted. Hence, the best match between a shapelet and a time series is searched for anywhere in the time series. We observed that for many data sets from the field of spectroscopy, time series from the same class show similar behavior patterns at similar locations along the frequency axis. Fig. 1 presents examples of two data sets which strongly support this insight. Based on this insight, we propose a new property as part of the definition of a shapelet, derived from the location in the time series from which the shapelet was extracted. This property limits the scope of the search for the best match of a shapelet to a time series to the vicinity of the location from which the shapelet was extracted. The assumption of locality is justified as spectroscopic measurements of items with similar properties should have very similar spectroscopic fingerprints, especially in areas characteristic of a specimen which are not expected to be contaminated. Our current implementation assumes that all time series are of equal length.

Although the time series are generally aligned, some allowance for misalignment is necessary. We therefore introduce a method for learning the misalignment characteristic of a data set. We evaluate our approach on data sets from the field of spectroscopy, and show that local-shapelets can reduce learning and classification time (especially for data sets with long time series) by over two orders of magnitude without impairing accuracy. For reproducibility, we have made all our code available online (Local shapelets, 2014).

The rest of the article is organized as follows: First we present basic definitions required for understanding the article and shortly describe the YK-algorithm (Section 2). Then we present related work (Section 3), followed by a description of local-shapelets and our proposed method for determining the range to examine (Section 4). Next we present our experimental evaluation (Section 5) followed by a brief statistical analysis, which provides a theoretical justification for our local-shapelets approach (Section 6). Finally, we summarize our results and present additional research directions to pursue (Section 7).

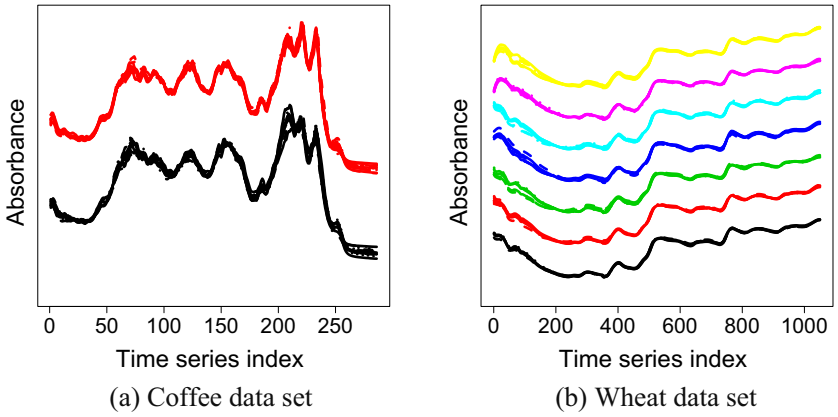
## 2. Background

Here we present a number of definitions necessary for the proper understanding of this article and a short description of the original shapelet algorithm as it is the basis of our work.

### 2.1. Definitions

**Definition 1.** A time series $T$ of length $m$ is a series of $m$ consecutive equally spaced measurements:

$T = t_0, t_1, \ldots, t_{m-1}$.




(a) Coffee data set (b) Wheat data set

**Fig. 1.** Examples of two data sets (coffee and wheat) from the field of spectroscopy. The time series of each class are vertically separated and in different colors. As shown, the examples of each class are tightly aligned, i.e., similar patterns are exhibited at similar locations along the *x*-axis.