Pattern Recognition Letters 45 (2014) 99–105

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Using derivatives in a longest common subsequence dissimilarity measure for time series classification ☆

Tomasz Górecki *

Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland

ARTICLE INFO

Article history:
Received 12 November 2013
Available online 27 March 2014

Keywords:
Longest common subsequence
Derivative longest common subsequence
Data mining
Time series

ABSTRACT

Over recent years the popularity of time series has soared. Given the widespread use of modern information technology, a large number of time series may be collected. As a consequence there has been a dramatic increase in the amount of interest in querying and mining such data. A vital component in many types of time series analyses is the choice of an appropriate dissimilarity measure. Numerous measures have been proposed to date, with the most successful ones based on dynamic programming. One of such measures is longest common subsequence (LCSS). In this paper, we propose a parametrical extension of LCSS based on derivatives. In contrast to well-known measures from the literature, our approach considers the general shape of a time series rather than point-to-point function comparison. The new dissimilarity measure is used in classification with the nearest neighbor rule. In order to provide a comprehensive comparison, we conducted a set of experiments, testing effectiveness on 47 real time series. Experiments show that our method provides a higher quality of classification compared with LCSS on examined data sets.

1. Introduction

Time series classification has been studied extensively by the machine learning and data mining communities. Such series are suitable for representing social, economic and natural phenomena, medical observations, and results of scientific and engineering experiments. The crucial point in time series classification is how to measure the dissimilarity of time series (a very good overview of dissimilarity measures can be found in [9]). The simplicity and efficiency of Euclidean distance [11] makes this the most popular dissimilarity measure in time series data mining [1,18]. It requires that both input sequences be of the same length, and it is sensitive to distortions and shifting along the time axis [29,25]. Such problems can be handled by elastic dissimilarity measures such as Dynamic Time Warping (DTW) [5] and Longest Common SubSequence (LCSS) [2,28]. DTW searches for the best alignment between two time series, attempting to minimize the distance between them. LCSS finds the length of the longest matching subsequence. Compared with Euclidean distance, DTW and LCSS are more elastic, supporting local time shifts and variations in lengths of pairs of time series, but they are also more expensive to compute. Of the three measures, LCSS is the least sensitive to noise, because it includes a threshold to define a “match” [28].

The effectiveness of the nearest neighbor classifier depends on the dissimilarity measure used to compare objects in the classification process. At present, the dissimilarity functions used in time series classification mostly involve point-to-point comparison of time series. The measures often reduce such distortions as occur if two time series do not have the same length or are locally out of phase, etc. It seems that in the classification domain there could be objects for which function value comparison is not sufficient. There could be cases where assignment to one of the classes depends on the general shape of objects rather than on strict function value comparison. An object associated with a function that responds to its variability in “time” is the derivative of the function. The function’s derivative determines areas where the function is constant, increases or decreases, and the intensity of the changes. The derivative determines the general shape of the function rather than the value of the function at a particular point. The derivative shows what happens in the neighborhood of the point. While the first derivative gives some information about the shape of the function (increasing or decreasing), the second derivative adds additional information as to where the function is convex or concave. We cannot expect that it will be sufficient to compare only time series derivatives. It seems that the best approach is to create a method which considers both the function values of time series

☆ This paper has been recommended for acceptance by A. Marcelli.

* Tel.: +48 618295308.
E-mail address: tomasz.gorecki@amu.edu.pl

http://dx.doi.org/10.1016/j.patrec.2014.03.009
0167-8655/© 2014 Elsevier B.V. All rights reserved.

and values of the derivative (or derivatives) of the function (shape comparison). The intensity of the influence of these approaches should be parameterized. Then we can expect that for different time series the method will select the appropriate intensities of these comparisons and give the best classification results.

In this paper we construct a dissimilarity measure that considers the above-mentioned approaches to time series classification. Consequently we are able to deal with situations where the investigated sequences are not different enough. For a dissimilarity function, a new parameterized family of dissimilarity measures is formed, where a fixed dissimilarity measure is used to compute dissimilarities of time series (function values) and their variability in "time" (dissimilarities of their derivatives). The new dissimilarity functions so constructed are used in the nearest neighbor classification method. The use of derivatives in time series classification is not a novelty. Some ideas of dissimilarity between trajectories using derivatives were proposed by Kosmelj [20] and Carlier [6]. They used the concepts of velocity and acceleration to measure the dissimilarities between trajectories in cluster analysis. D'Urso and Vichi [10] and Coppi et al. [7] developed this idea and used it to perform cluster analysis of longitudinal data. The use of derivatives with DTW was proposed by Keogh and Pazzani [17]. However they used only the dissimilarity between the derivatives, rather than the standard dissimilarity between the time series. Górecki [13] and Górecki and Łuczak [14] presented results concerning derivative DTW where just the first derivative is added, while parameterization involves both the function and derivative. Such an approach was shown to give very good results. Górecki and Łuczak [15] also presented results where the second derivative is added. The parametric approach makes it possible to adapt to the data set, but without overtraining. Now we try to extend this methodology to LCSS, which is a better method than DTW in the presence of outliers [28] and generally is very close to the best dissimilarity measure DTW [21].

In this paper we first review the concept of time series and the longest common subsequence dissimilarity measure (Section 2). At the end of that section we introduce our dissimilarity measure based on derivatives. The data sets used and the experimental setup are described in Section 3. Section 4 contains the results of our experiments on the described real data sets, as well as statistical analysis of the results and analysis of the running times of the investigated methods. Conclusions are given in Section 5.

2. Methods

2.1. Longest common subsequence

The longest common subsequence dissimilarity measure is a variation of the edit dissimilarity measure used in speech recognition. The basic idea is to match two sequences by allowing them to stretch, without rearranging the sequence of the elements but allowing some elements to be unmatched or left out (e.g., outliers) – whereas in Euclidean Distance and DTW, all elements from both sequences must be used, even the outliers. The overall idea is to count the number of pairs of points from the two sequences that match. One point can never be associated twice with points of the other sequence, so that the maximum number of associations is the smaller length of the two sequences. The LCSS measure has two parameters, δ and ε (Fig. 1). The constant δ, which is usually set to a percentage of the sequence length, is a warping threshold and controls the window size for matching a given point from one sequence to a point in another sequence. It controls how far in time we can go in order to match a given point from one trajectory to a point in another trajectory. The constant $0 < \varepsilon < 1$ is the matching threshold: two points from two sequences are considered to match if their distance is less than ε.

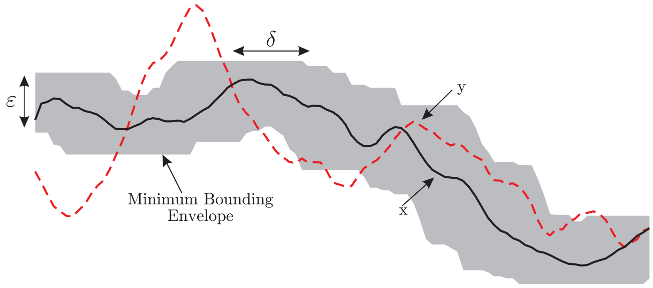

Fig. 1. Matching within δ in time and ε in space. Everything outside the bounding envelope can never be matched.

Longest common subsequences of the time series x and y of length n and m may be recursively defined as follows:

$$L(i,j) = \begin{cases} 0 & \text{for } i = 0 \\ 0 & \text{for } j = 0 \\ 1 + L[i-1, j-1] & \text{for } |x_i - y_j| < \varepsilon \text{ and } |i-j| \leqslant \delta \\ \max(L[i-1,j], L[i,j-1]) & \text{in other cases} \end{cases}$$

$L(n,m)$ contains the similarity between x and y, because it corresponds to the length of the longest common subsequence of elements between time series x and y. To define the dissimilarity between x and y, we can compute [24]:

$$\mathrm{LCSS}(x,y) = \frac{n+m-2L(n,m)}{n+m}. \quad (1)$$

According to this definition, this measure takes values from $(1 - 2\frac{\min(n,m)}{n+m})$ to 1. For two trajectories of equal length it takes values from 0 to 1.

Taking into account only sufficiently similar points, LCSS solves the problem of the presence of noise, but does not satisfy the triangle inequality [28], so it is not a distance metric. LCSS is robust to noise and is expected to be more accurate than DTW in the presence of outliers.

2.2. Longest common subsequence based on derivatives

Let LCSS be the longest common subsequence dissimilarity measure for two time series x and y.

2.2.1. 2D method

A dissimilarity measure which considers both the function values of time series and values of the first derivative is defined by:

$$\mathrm{DD}_{\mathrm{LCSS}}(x,y) := a\,\mathrm{LCSS}(x,y) + b\,\mathrm{LCSS}(\nabla x, \nabla y), \quad (2)$$

where ∇x and ∇y are the first discrete derivatives of x, y, and $a, b \in [0,1]$ are parameters. The discrete derivative of a time series x with length n is defined by

$$\nabla x(i) = x(i+1) - x(i), \quad i = 1, 2, \ldots, n-1. \quad (3)$$

2.2.2. 3D method

A dissimilarity measure which considers both the function values of time series and values of the first and second derivatives is defined by:

$$2\mathrm{DD}_{\mathrm{LCSS}}(x,y) := a\mathrm{LCSS}(x,y) + b\mathrm{LCSS}(\nabla x, \nabla y) + c\mathrm{LCSS}(\nabla^2 x, \nabla^2 y), \quad (4)$$

where $\nabla^2 x$ and $\nabla^2 y$ are the second discrete derivatives of x, y, and $a, b, c \in [0,1]$ are parameters.

If the similarity measure in the above definitions is a metric, then the new measures $\mathrm{DD}_{\mathrm{LCSS}}$ and $2\mathrm{DD}_{\mathrm{LCSS}}$ are also metrics. In