Msc Basismodul

Report: Similarity Search in a Time Series Database

Dominik Frey
University of Zurich
Department of Informatics
Supervisor: Kevin Wellenzohn

1. Introduction

This report aims to describe a small project which was undertaken to describe the behavior of three different algorithms, namely the Euclidean Distance algorithm, the Dynamic Time Warping algorithm (DTW) and the Dynamic Time Warping algorithm using the Sakoe-Chiba band, in terms of their time and space complexity. All these three algorithms can be used in order to describe the similarity of two different time series. A time series is a sequence of data points collected over a specific time interval. The amount of collected data can get very large and therefore algorithms which are able to process such large amounts of data efficiently are needed.

During the project the time and space complexity of the algorithms were analyzed. To verify the theoretical results, we implemented a small program that allows the user to query for the most similar time series in an archive with all the different algorithms. To test the implementation the UCR Time Series Classification Archive was used [1].

The first part of the report is dedicated to the functionality of the algorithms. Firstly, the Euclidean distance and the DTW algorithms are described and secondly the DTW with the Sakoe-Chiba band is introduced so as to show an opportunity to optimize the runtime complexity of the DTW algorithm. In the second part the results of the analysis of the algorithms are presented.

2. Functionality of the algorithms

The Euclidean Distance and the DTW are different in the core of their functionality whereas the DTW with the Sakoe-Chiba band is a modification of the regular Dynamic Time Warping algorithm.

2.1. Euclidean Distance

The Euclidean Distance is one of the most popular ways to describe the similarity of two different time series. A disadvantage of this algorithm is that the time series have to be of the same length. This is different with DTW as we will see later on. Let $|s_1|$ be the length of time series s_1 and s_{1t} the value of the same time series at time t. Then the Euclidean Distance is expressed as:

$$ED(s_1, s_2) = \sqrt{\sum_{t=1}^{|s_1|} (s_{1t} - s_{2t})^2}$$

This algorithm operates on every single pair of data points of the two time series which occur at the same point in time.

2.2 Dynamic Time Warping

The second very popular algorithm in the field of similarity search in time series database is Dynamic Time Warping. Originally the algorithm was developed to quantify the similarity of spoken words in speech recognition [2].

One of the advantages of DTW is that we can compare time series which are not of the same length. But the more important point is that DTW allows to compare time series which are out of phase and that these phases can be of different length. In the example of speech recognition this means that we can compute the similarity of two spoken words which are spoken at different speed.

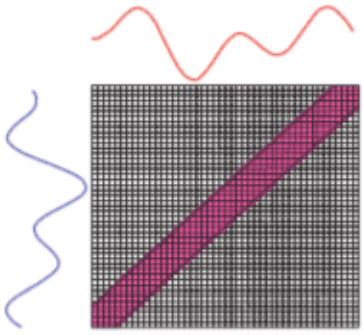
Suppose we have two time series S_1 and S_2 of length m and n respectively where:

$$S_1 = s_{11}, \ldots, s_{1i}, \ldots, s_{1m};$$
$$S_2 = s_{21}, \ldots, s_{2i}, \ldots, s_{2n};$$

Then we can construct a m-by-n matrix, where m and n are the length of the respective time series, and align the two sequences S_1 and S_2. Each element (i^{th}, j^{th}) of the matrix contains the squared distance $d(s_{1i}, s_{2j}) = (s_{1i} - s_{2j})^2$ between the two points s_{1i} and s_{2i} [2]. Then the minimal distance between the two time series is calculated according to the dynamic programming algorithm, meaning that in a cell of the matrix is not only a single distance anymore, but the cost which is the cumulative distance to reach this cell. This allows to compare two similar time series which are out of phase.

2.3 DTW with Sakoe-Chiba band

The Sakoe-Chiba band was developed as an optimized version of the DTW in order to reduce time complexity. To illustrate the concept of this algorithm a new constant called warping window w is introduce. This constant works as a constraint to the dynamic programming algorithm in a way, that the algorithm does not go through the whole sequence Q but only within the range of $\pm w$ along the diagonal of the matrix, as long as it works within the borders of the matrix. The width of this constraint, the warping window, is generally set to be 10% of the length of the time series Q [3]. The algorithm is illustrated in Picture 1 [3].



Picture 1

As the reader might recognize now, the Euclidean Distance is a special case of DTW in a sense that its constraint is limited to only the diagonal of the matrix with a warping window of $w = \pm 0$.

3. Findings

As already mentioned in the introduction the purpose of this project was to analyze the runtime and space complexity of the above described algorithms on the basis of the implemented program and the UCR Time Series Classification Archive. Obviously, the Euclidean Distance is the fastest one and needs linear time $O(n)$ to go through two time series of the same length n. Therefore to operate on the sequences, space of the size of n needs to be allocated. Thus, the space complexity is $O(n)$.

For the DTW algorithm we first need to allocate a matrix of the size m x n with m and n again as the length of the respective time series. This matrix is then filled with the distance values between the (i^{th}, j^{th}) elements of the respective time series. As mentioned above in the example of two time series S_1 and S_2 this would be $d(s_{1i}, s_{2i})$ for the respective entry in the matrix. Therefore the space complexity of DTW is $O(mn)$. To calculate the time complexity we need to take into account that DTW operates on a nested for-loop in order to get the distance values. When we assume two time series of lengths m and n respectively, we can derive that DTW takes $O(mn)$ time.

The analysis of the Sakoe-Chiba band constraint yielded a space complexity of $O(mn)$. Even though the algorithm does not go through all the elements of the inner for-loop, which operates on the time series S_2 , we still need to keep the whole matrix in the RAM. So space of the size m x n needs to be allocated which leeds to a space complexity of $O(mn)$. As said before, we do not need to go through the whole inner for-loop in each step of the dynamic programming algorithm, but only a fraction of it. This fraction is represented by the window parameter w . Hence the time complexity of DTW with a Sakoe-Chiba band is $O(mw)$, which is less than $O(mn)$ from the regular DTW algorithm. The findings are summarized in Table 1.

	Euclidean Distance	DTW	DTW with Sakoe-Chiba band
Time Complexity	$O(n)$	$O(mn)$	$O(mn)$
Space Complexity	$O(n)$	$O(mn)$	$O(mw)$

Table 1

References

[1] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/ time_series_data/.

[2] J. B. Kruskal and M. Liberman. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison*, chapter 4. CSLI Publications, Stanford,

CA 94305, 1983.

[3] V. Niennattrakul and C. A. Ratanamahatana, Learning DTW Global Constraint for Time Series Classification, CoRR abs/0903.0041 (2009)