The Influence of Global Constraints on DTW and LCS Similarity Measures for Time-Series Databases

Vladimir Kurbalija[1], Miloš Radovanović[1], Zoltan Geler[2], and Mirjana Ivanović[1]

[1] Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Trg D. Obradovica 4, 21000 Novi Sad, Serbia, {kurba,radacha,mira}@dmi.uns.ac.rs
[2] Faculty of Philosophy, University of Novi Sad, Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia
gellerz@gmail.com

Abstract. Analysis of time series represents an important tool in many application areas. A vital component in many types of time-series analysis is the choice of an appropriate distance/similarity measure. Numerous measures have been proposed to date, with the most successful ones based on dynamic programming. Being of quadratic time complexity, however, global constraints are often employed to limit the search space in the matrix during the dynamic programming procedure, in order to speed up computation. In this paper, we investigate two representative time-series distance/similarity measures based on dynamic programming, Dynamic Time Warping (DTW) and Longest Common Subsequence (LCS), and the effects of global constraints on them. Through extensive experiments on a large number of time-series data sets, we demonstrate how global constrains can significantly reduce the computation time of DTW and LCS. We also show that, if the constraint parameter is tight enough (less than 10–15% of time-series length), the constrained measure becomes significantly different from its unconstrained counterpart, in the sense of producing qualitatively different 1-nearest neighbour (1NN) graphs. This observation highlights the need for careful tuning of constraint parameters in order to achieve a good trade-off between speed and accuracy.

Keywords: Time series, DTW, LCS, Global Constraints

1 Introduction

In many scientific fields, a time series consists of a sequence of values or events obtained over repeated measurements of time [1]. Time-series analysis is comprised of methods that attempt to understand time series, to explain the underlying context of the data points or to make forecasts.

Time-series databases are popular in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, process and quality control, observation of natural phenomena, scientific and engineering experiments,

medical treatments, etc. As a consequence, the last decade witnessed an increasing interest in querying and mining such data, which resulted in a large amount of work introducing new methodologies for different task types including: indexing, classification, clustering, prediction, segmentation, anomaly detection, etc. [1, 2, 3, 4]

One of the most important aspects of time-series analysis is the choice of appropriate similarity/distance measure – the measure which tells to what extent two time series are similar. However, unlike data types in traditional databases where the similarity/distance definition is straightforward, the distance between time series needs to be carefully defined in order to reflect the underlying (dis)similarity of these specific data, which is usually based on shapes and patterns. As expected, there exists a large number of measures for expressing (dis)similarity of time-series data proposed in the literature, e.g., Euclidean distance (ED) [2], Dynamic Time Warping (DTW) [5], distance based on Longest Common Subsequence (LCS) [6], Edit Distance with Real Penalty (ERP) [7], Edit Distance on Real sequence (EDR) [8], Sequence Weighted Alignment model (Swale) [9].

Many of these similarity measures are based on dynamic programming. It is well known that the computational complexity of dynamic programming algorithms is quadratic, which is often not suitable for larger real-world problems. However, the usage of global constraints such as Sakoe-Chiba band [21] and Itakura parallelogram [22] can significantly speed up the calculation of similarities. Furthermore, it is also reported [10] that the usage of global constraints can improve the accuracy of classification compared to unconstrained similarity measures. The accuracy of classification is commonly used as a qualitative assessment of a similarity measure.

In this paper we will investigate the influence of global constraints on two most representative similarity measures for time series based on dynamic programming: DTW and LCS. We will report the calculation times for different sizes of constraints in order to explore the speed-up gained from these constraints. Also, the change of the 1-nearest neighbour graph will be explored with respect to the change of the constraint size. The proposed research will provide a better understanding of global constraints and offer deeper insight into their advantages and limitations.

All experiments presented in this paper are performed using the system FAP (Framework for Analysis and Prediction) [11]. The data for experiments is provided by the UCR Time Series Repository [12], which includes the majority of all publicly available, labelled time-series data sets in the world.

The rest of the paper is organized as follows. The next section presents the necessary background knowledge about similarity measures and gives an overview of related work. Section 3 briefly describes the FAP system used for performing experiments. The methodology and results of extensive experiments are given in Section 4. Section 5 concludes the paper and presents the directions for further work.

2 Background and Related Work

The Euclidean metric is probably the most intuitive metric for time series, and as a consequence very commonly used [2, 13, 14, 15, 16]. In addition, it is also very fast –

its computation complexity is linear. The distance between two time series is calculated as a sum of distances between corresponding points of two time series. However, it became evident that this measure is very brittle and sensitive to small translations across the time axis [10, 17].

Dynamic Time Warping (DTW) can be considered as a generalization of Euclidian distance where it is not necessary that the *i*-th point of one time series must be aligned to the *i*-th point of the other time series [10, 17, 18, 19]. This method allows elastic shifting of the time axis where in some points time "warps". The DTW algorithm computes the distance by finding an optimal path in matrix of distances between points of two time series. The Euclidian distance can be seen as special case of DTW where only the elements on the main diagonal of the matrix are taken into account.

Longest Common Subsequence (LCS) applies a different methodology. According to LCS, the similarity between two time series is expressed as the length of the longest common subsequence of both time series [20].

Both DTW and LCS are based on dynamic programming – the algorithms seek the optimal path in the search matrix. The types of matrices are different but the approach is the same. DTW examines the matrix of distances between points, while LCS examines the matrix of longest common subsequences of different-length subseries. As a consequence, both algorithms are quadratic. However, the introduction of global constraints can significantly improve the performance of these algorithms. Global constraints narrow the search path in the matrix, which results in a significant decrease in the number of performed calculations. The most frequently used global constraints are the Sakoe-Chiba band [21] and the Itakura parallelogram [22]. These constraints were introduced to prevent some bad alignments, where a relatively small part of one time series maps onto a large section of another time series.

The quality of similarity measures is usually evaluated indirectly, e.g. by assessment of classifier accuracy. The simple method combining the 1NN classifier and some form of DTW distance was shown to be one of the best-performing time-series classification techniques [4, 17, 18, 23]. In addition, the accuracy of 1NN directly reflects the quality of a similarity measure. Therefore, in this paper we report the calculation times for unconstrained and constrained DTW and LCS, and we focus on the 1NN graph and its change with regard to the change of constraints. The influence of global constraints is not investigated well in the literature, and the results presented in this paper will provide a better understanding of theirs essence.

3 The FAP System

There are three important concepts which need to be considered when dealing with time series: pre-processing transformation, time-series representation and similarity measure. The task of pre-processing transformations is to remove different kinds of distortions in raw time series. The task of time-series representation is to reduce the usually very high dimensionality of time series while preserving their important properties. Finally, the task of a similarity measure is to reflect the essential similarity of time series, which are usually based on shapes and patterns.

All these concepts, when introduced, are usually separately implemented and presented in different publications. Every newly-introduced representation method or distance measure has claimed a particular superiority [4]. However, this was usually based on comparison with only a few other representatives of the proposed concept. On the other hand, to the best of our knowledge there is no freely available system for time-series analysis and mining which supports all mentioned concepts, with the exception of the work proposed in [4]. Being motivated by these observations, we have designed a multipurpose, multifunctional system FAP – Framework for Analysis and Prediction [11]. FAP supports all mentioned concepts: representations, similarity measures and pre-processing tasks; with the possibility to easily change some existing or to add new concrete implementation of any concept.

At this stage of development, all main similarity measures (L_p, DTW, CDTW (Constrained DTW), LCS, CLCS, ERP, CERP, EDR, CEDR and Swale) are implemented, and the modelling and implementation of representation techniques is in progress. All constrained measures employ the Sakoe-Chiba band. Furthermore, several classifiers and statistical tests are also implemented.

4 Experimental Evaluation

In this section we will investigate the influence of global constraints on two most illustrative similarity measures based on dynamic programming: DTW and LCS. Furthermore, two aspects of applying global constraints are considered: efficiency and effectiveness of the 1NN classifier for different values of constraints. For both similarity measures, the experiments are performed with the unconstrained measure and a measure with the following constraints: 75%, 50%, 25%, 20%, 15%, 10%, 5%, 1% and 0% of the size of the time series. This distribution was chosen because it is expected that measures with larger constraints behave similarly to the unconstrained measure, while smaller constraints exhibit more interesting behaviour [10, 18].

A comprehensive set of experiments was conducted on 38 data sets from [12], which includes the majority of all publicly available, labelled time-series data sets currently available for research purposes. The length of time series varies from 24 to 1882 depending on the data set. The number of time series per data set varies from 60 to 9236.

4.1 Computational Times

In the first experimental phase we wanted to investigate the influence of global constraints on the efficiency of calculating the distance matrix. The distance matrix for one data set is the matrix where element (i,j) contains the distance between *i*-th and *j*-th time series from the data set. The calculation of the distance matrix is a time-consuming operation, which makes it suitable for measuring the efficiency of global constraints.

In Table 1, the calculation times of DTW in milliseconds are given for some datasets and for different values of constraints. Table 2 contains the same data for the

LCS measure. The complete tables are available in extended version of the paper at Computing Research Repository – CoRR (http://arxiv.org/corr/home). All experiments are performed on AMD Phenom II X4 945 with 3GB RAM.

Table 1. Calculation times of distance matrix for DTW

Name of dataset	DTW unconstrained	75%	50%	25%	20%	15%	10%	5%	1%	0%
Car	79844	73391	58656	34500	28562	22047	15141	8016	2016	672
CBF	258375	242703	198969	124719	105375	86031	62672	41766	23047	17203
cinc_ecg_torso	88609875	79638047	63711094	36991468	30533875	23531718	16107266	8290062	1814203	146672
fish	434297	392656	317093	185672	154672	119906	83672	45390	12031	3969
Haptics	4257391	3789922	3052828	1774359	1468031	1135609	781547	404907	88234	10844
Inlineskate	25407250	22014203	17571907	10359203	8563921	6618313	4534843	2341546	491640	32109
Lighting2	101641	90781	72750	42828	35171	27391	19890	9875	2594	719
Mallat	97847485	88454641	70498062	41180453	34403500	26572297	18149719	9492531	2189141	284188
OSULeaf	592328	536562	431062	254844	210047	164515	113000	61312	17672	5906

Table 2. Calculation times of distance matrix for LCS

Name of dataset	DTW unconstrained	75%	50%	25%	20%	15%	10%	5%	1%	0%
Car	52282	48844	39187	23078	18984	14875	10281	5516	1531	609
CBF	170016	161250	134297	85671	72359	60656	45797	32406	20204	17547
cinc_ecg_torso	62996109	59056579	46699297	27158750	22369187	17262734	11882062	6155704	1402765	161766
fish	285922	268282	214875	127484	105765	82329	58000	31906	9390	4156
Haptics	2793547	2593672	2074313	1218812	1008203	777266	539672	280437	64437	12359
Inlineskate	16524687	15507859	12291281	7163172	5895296	4556593	3139500	1633141	351640	37016
Lighting2	71000	66578	53329	31531	26422	20484	14188	7547	2250	750
Mallat	67029406	62582938	50095359	29302062	24642968	18588750	12998219	6910859	1670969	307656
OSULeaf	388375	371156	295813	174734	144938	114547	79140	43750	13984	6375

It is evident that the introduction of global constraints in both measures significantly speeds up the process of distance matrix computation, which is the direct consequence of a faster similarity measure. The difference of computation times between an unconstrained measure and a measure with a small constraint is two and somewhere three orders of magnitude. Furthermore, it is known for DTW that smaller values of constraints can tend to more accurate classification [10]. The authors also reported that the average constraint size, which gives the best accuracy, for all datasets is 4% of the time-series length. On the other hand, the influence of global constraints on the LCS measure is still not well investigated. However, it is evident that the usage of global constraints contributes to the efficiency of both measures, and, at least for DTW, improves classification accuracy.

4.2 The Change of 1NN Graph

In the next experimental phase we wanted to investigate the influence of global constraints on the NN graph of each dataset. This decision was mainly motivated by the fact that the 1NN classifier is among the best classifiers for time series [18].

The nearest neighbour graph is a directed graph where each time series is connected with its nearest neighbour. We calculated this graph for unconstrained measures (DTW and LCS) and for measures with the following constraints: 75%, 50%, 25%, 20%, 15%, 10%, 5%, 1% and 0% of the length of time series. After that,

we focused on the change of the 1NN graph for different constraints compared to the graph of the unconstrained measure. The change of nearest-neighbour graphs is tracked as the percentage of time series (nodes in the graph) that changed their nearest neighbour compared to the nearest neighbour according to the unconstrained measure. The graphical representation of results can be seen in Figure 1 and Figure 2 for DTW and LCS, respectively. Each figure is represented by two charts showing one half of the data sets for the sake of readability. The numerical results are available at CoRR.

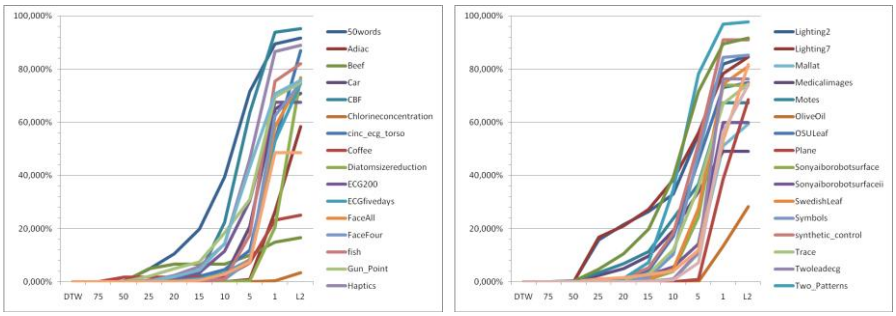

Fig. 1. Change of 1NN graph for DTW

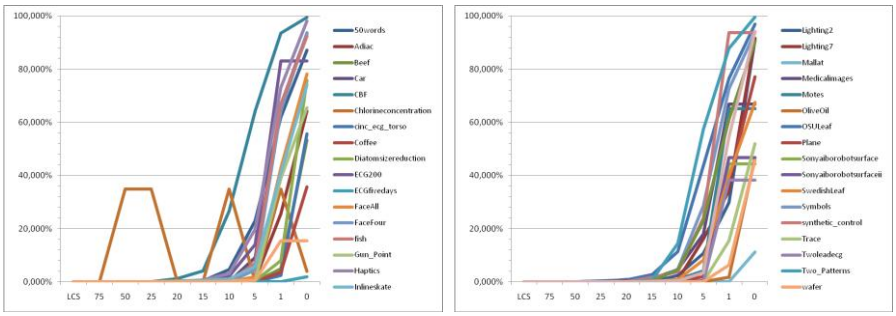

Fig. 2. Change of 1NN graph for LCS

The presented results clearly show that both measures behave in a similar manner when the constraint is narrowed. The 1NN graph of the DTW measure remains the same until the size of the constraint is narrowed to approximately 50%, and after that the graph starts to change significantly. The situation with LCS is more pronounced: the LCS 1NN graph remains the same to approximately 10-15%, while for smaller constraints it changes even more drastically.

Only one data set does not follow this rule for LCS measure: *Chlorineconcentration*. For some values of the constraint (75%, 20%, 15% and 5%) the graph is the same as the unconstrained, while for other values of the constraint the difference of graphs is 34.87%. Additionally, we investigated the structure of this dataset and found that the time series are periodical, where all time series have approximately the same period. Since the LCS measure searches for the longest common subsequence, it turns out that for some constraint values the LCS algorithm

finds the same sequence as the unconstrained LCS. Other values of the constraint break that sequence, which is then no more longest, and as a consequence some other time series is found as a nearest neighbour. This behaviour is caused by the strict periodicity of this data set.

All other datasets (for both measures) reach high percentages of difference (over 50%) for small constraint sizes (5-10%). This means that when the constraint size is narrowed to 10% of the length of time series, then more than 50% of time series in the data set change their first neighbour with regard to the unconstrained measure. This fact strongly suggests that constrained measures represent qualitatively different measures than the unconstrained ones.

5 Conclusions and Future Work

Although the Euclidian measure is simple and very intuitive for time-series data, it has a known weakness of sensitivity to distortion in the time axis. Many elastic measures (DTW, LCS, ERP, EDR, etc.) were proposed in order to overcome this weakness. However, they are all based on dynamic programming and have quadratic computation complexity. Global constraints are introduced in dynamic programming algorithms to narrow the search path in the matrix and to decrease computation time.

In this paper, we examined the influence of global constraints on two most representative elastic measures for time series: DTW and LCS. Through an extensive set of experiments, we showed that the usage of global constraints can significantly reduce the computation time of these measures, which is their main weakness. In addition, we demonstrated that the constrained measures are qualitatively different than their unconstrained counterparts. For DTW it is known that the constrained measures are more accurate than the unconstrained, while for LCS this issue is still open.

In future work we plan to investigate the accuracy of the constrained LCS measure for different values of constraints. It would also be interesting to explore the influence of global constraints on the computation time and 1NN graphs of other elastic measures like ERP, EDR, Swale, etc. Finally, the constrained variants of these elastic measures should also be tested with respect to classification accuracy.

Acknowledgments. The authors acknowledge the support of this work by the Serbian Ministry of Education and Science through project "Intelligent Techniques and Their Integration into Wide-Spectrum Decision Support" no. OI174023.

References

1. Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA, 2005.
2. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast Subsequence Matching in Time-Series Databases. In SIGMOD Conference, 1994, pp. 419-429.

3. E. J. Keogh. A Decade of Progress in Indexing and Mining Large Time Series Databases. In VLDB, 2006, pp. 1268-1268.
4. Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. In VLDB '08, 2008, Auckland, New Zealand, pp. 1542-1552.
5. E. J. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. Knowl. Inf. Syst., 7(3), 2005, pp. 358-386.
6. M. Vlachos, D. Gunopulos, and G. Kollios. Discovering similar multidimensional trajectories. In ICDE, 2002, pp. 673-684.
7. L. Chen and R. T. Ng. On the marriage of lp-norms and edit distance. In VLDB, 2004, pp. 792-803.
8. L. Chen, M. T. ¨Ozsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In SIGMOD Conference, 2005, pp. 491-502.
9. M. D. Morse and J. M. Patel. An efficient and accurate method for evaluating time series similarity. In SIGMOD Conference, 2007, pp. 569-580.
10. Ratanamahatana, C., Keogh. E. Three Myths about Dynamic Time Warping. In proc. of SIAM Inter. Conf. on Data Mining, Newport Beach, CA, April 2005, pp. 506-510.
11. V. Kurbalija, M. Radovanović, Z. Geler and M. Ivanović. A framework for time-series analysis. In Proceedings of the 14th international conference on Artificial intelligence: methodology, systems, and applications (AIMSA'10), 2010, pp. 42-51.
12. E. Keogh, X. Xi, L. Wei, and C. Ratanamahatana. The UCR Time Series Classification/Clustering Page: http://www.cs.ucr.edu/~eamonn/time_series_data/, 2006.
13. Agrawal R, Lin KI, Sawhney HS, Shim K, Fast similarity search in the presence of noise, scaling, and translation in times-series databases. In: Proceedings of the 21st international conference on very large databases, 1995, pp 490–501.
14. Chan KP, Fu A, Yu C, Haar wavelets for efficient similarity search of time-series: with and without time warping. IEEE Trans Knowl Data Eng 15(3), 2003, pp. 686–705.
15. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S, Dimensionality reduction for fast similarity search in large time series databases. J Knowl Inf Syst 3(3), 2000, pp. 263–286.
16. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S, Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of ACM SIGMOD conference on management of data, May 2001, pp 151–162.
17. Keogh, E. Exact indexing of dynamic time warping. In 28th International Conference on Very Large Data Bases. Hong Kong. 2002. pp 406-417.
18. Xi X, Keogh E, Shelton C, Wei L, and Ratanamahatana CA. Fast time series classification using numerosity reduction. In Proceedings of the 23rd international conference on Machine learning (ICML '06). ACM, New York, NY, USA, 2006, pp. 1033-1040.
19. Berndt D, Clifford J, Using dynamic time warping to find patterns in time series. AAAI-94 workshop on knowledge discovery in databases, 1994, pp 229–248.
20. M. Vlachos, G.Kollios, D. Gunopulos: Discovering Similar Multidimensional Trajectories, In Proc. of 18th Inter. Conf. on Data Engineering (ICDE), San Jose, CA, 2002, pp. 673-684.
21. Sakoe H, Chiba S, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoustics Speech Signal Process ASSP 26, 1978, pp. 43–49.
22. Itakura F, Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoustics Speech Signal Process ASSP 23, 1975, pp. 52–72.
23. M. Radovanović, A. Nanopoulos and M. Ivanović. Time-series classification in many intrinsic dimensions. In Proceedings of SDM'10, 10th SIAM International Conference on Data Mining, Columbus, Ohio, USA, 2010, pp. 677–688.