Large-Scale Similarity-Based Time Series Mining

Diego F. Silva[1] **Gustavo E. A. P. A. Batista (advisor)**[1] **, Eamonn Keogh (co-advisor)**[2]

[1] Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo

[2]Department of Computer Science and Engineering – University of California, Riverside

diegofsilva@usp.br, gbatista@icmc.usp.br, eamonn@cs.ucr.edu

Abstract. *Measuring the (dis)similarity between time series is the main procedure of several algorithms for mining this kind of data, which is ubiquitous in the day-by-day of human beings. While providing satisfactory results, similarity-based methods usually suffer from a high time complexity. This work summarizes a thesis on developing algorithms that allow the similarity-based mining of temporal data in a large scale. The contributions of the thesis have implications in several data mining tasks, such as classification, clustering and motif discovery, as well as applications in music data science.*

1. Introduction

Time series are ubiquitous in the day-by-day of human beings. A diversity of application domains such as medicine, biology, economics, and signal processing generate data arranged in time. Consequently, the time series analysis has attracted the attention and effort of many researchers around the world. Many methods for analyzing this kind of data have been proposed for different temporal data mining tasks, such as classification, clustering, motif discovery, and anomaly detection. Several of these methods are constructed over the (dis)similarity relations between the time series objects in the dataset.

Dynamic Time Warping (DTW) is arguably the most relevant distance measure for time series analysis. Such relevance has been evidenced, for instance, by a large body of experimental research on classification and clustering. Also, using DTW is pointed as future direction in some other tasks, like in the motif discovery. The main drawback of DTW is its computational complexity. The algorithm to calculate DTW is a dynamic programming technique that requires a quadratic matrix concerning the length of the time series. Although we can use a simple trick to reduce it to a linear space complexity, there is no known exact algorithm to reduce its time complexity.

To avoid this issue, a common approach is applying a less costly distance measure, such as the Euclidean distance (ED). However, the ED is very sensitive to small distortions in the time axis, as demonstrated by the alignments presented by Figure 1. While the ED is not able to correctly align slightly displaced features (peaks and valleys), the non-linear alignment obtained by DTW is robust, for instance, to the warped valley. Alternatively, if we are interested in applying DTW with mining algorithms that require the distance between plenty of pairs of time series, the only speed up techniques are approximations. However, these methods does not guarantee a limit for the approximation error, which may cause distortions on the space of distances.

In this thesis, we proposed the first method to speed up the exact DTW calculation that fits any distance-based time series mining algorithm. Also, we proposed a subtle

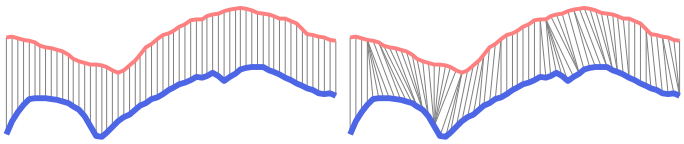

Figure 1. Alignment obtained by matching two subsequences by the ED (*left*) and the DTW (*right*)

modification of the DTW to provide invariance to an unnoticed distortion which is usual in time series streaming data. Finally, we also adapted strategies of similarity-based time series mining to deal with music data. All the proposed methods are more efficient and effective than the state-of-the-art algorithms.

2. Speeding Up the DTW Algorithm

The dynamic programming-based DTW's algorithm fills a two-dimensional cost matrix. However, the final distance value is given by a function of the values in a limited number of cells in this matrix. Exploring this fact, we developed PrunedDTW, the first *task-independent* algorithm to speed up the *exact* DTW calculation [Silva and Batista 2016]. Our method prunes the calculation of many cells that do not belong to the optimal path, and, consequently, speed up the distance computation. Figure 2 exemplifies the proportion between calculated and pruned cells when PrunedDTW is applied.

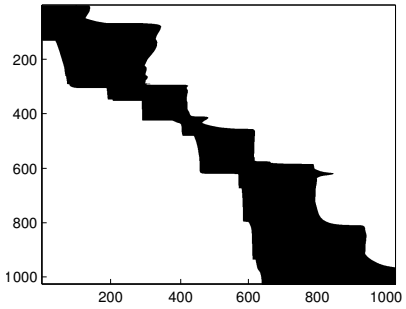

Figure 2. Regions of the DTW matrix pruned by our method (in white)

Initially, we focused our evaluation at the calculation of the all-pairwise distance matrix. Although this is a common requirement in a diversity of data mining applications, there was no exact methods to improve its time performance. In this scenario, our method performed up to ten times faster than the traditional DTW. Specifically, our method is more effective in the worst cases for DTW. In other words, the slower is the original algorithm, the higher is the speedup achieved by PrunedDTW.

Additionally, we successfully embedded PrunedDTW into the subsequence similarity search task, which is probably the most used operation in time series mining. For this reason, there is a plethora of algorithms to index the search and speed up this procedure. Rakthanmanon et al. [2012] compiled these contributions in the UCR Suite, considered the state-of-the-art tool for subsequence similarity search. According to the authors, the runtimes achieved by this suite are "close to the optimal."

We experimented the UCR Suite in different scenarios and verified that the distance calculations, even in a really reduced number, still are a tight bottleneck for the

search runtime. Considering this, we adapted the PrunedDTW to embed it into the UCR Suite [Silva et al. *in press*]. In this scenario, we can use statistics calculated in the indexing phase to further improve the pruning power of our method. We empirically demonstrated that, depending on the input parameters, our method performs between two and five times faster than the state-of-the-art. As in the all-pairwise scenario, the worst is the case, the higher is the speedup caused by PrunedDTW.

3. Improving DTW in the Streaming Time Series Mining

While many time series benchmark datasets guarantee a perfect segmentation of subsequences, most practical problems on time series mining require the automatic segmentation of time series in a streaming fashion. In this case, the presence of observations in the endpoints of the segmented subsequences which do not belong to the event of interest, which we call prefix and suffix, is virtually unavoidable. Although DTW is invariant to several aspects, it is very sensitive to spurious endpoints. Prefixes and suffixes may drastically change the DTW distance relation between the subsequences in a dataset, causing a significant decreasing of performance of the mining algorithm.

To circumvent this problem, we proposed the Prefix and Suffix Invariant DTW (ψ-DTW) [Silva et al. 2016a]. This distance measure is based on a subtle modification of the DTW's endpoints constraint, allowing the algorithm to skip the matching of some observations in the extremities of the time series. Figure 3 illustrates how the proposed algorithm matches the observations of the time series introduced by Figure 1.

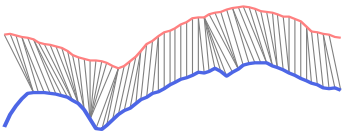

Figure 3. Alignment of time series points obtained by ψ-DTW

To test the robustness of our method, we compare its performance against the classification accuracy obtained by the classic DTW varied applications which generate streaming time series. We stated that ψ-DTW outperforms DTW in the classification in the presence of prefixes and suffixes with statistical significance.

While we demonstrated the effectiveness of ψ-DTW in time series classification, spurious endpoints are commonly presented in different temporal mining tasks. Particularly, motifs (reoccurring patterns) and discords (anomalous patterns) require a massive number of distance calculations in streaming time series data, being highly affected by these distortions. In an under-review paper, we show that using ψ-DTW avoids false positives and false dismissals commonly observed by applying other distance measures, improving the time series motifs and discords discovery in several application domains. Our algorithm is based on the Matrix Profile (MP) [Yeh et al. 2016], a novel representation of subsequence similarities in a long/streaming time series to find, in a single low memory-cost structure, time series motifs, discords, shapelets (representative subsequences of distinct classes), among other primitives.

As the original proposal of MP relies solely on the Euclidean distance, we proposed the first method able to construct such structure under a non-linear alignment distance. Specifically, we created a tool based on the UCR Suite adapted to ψ-DTW to allow

us to scale the motifs and discords discovery to the order of thousands of hundreds data points. The proposed method is one order of magnitude faster than a straightforward implementation of a DTW-based MP. Moreover, our method is anytime, which means that it can be interrupted at any moment and returns a good approximate answer. We empirically demonstrated that our method is able to return most of the best motifs (8 out of top 10 in random walk data) using fewer time than the required to calculate the ED-based MP.

4. Applications on Music Data Science

Although the main focus of this thesis is the time series mining, we expanded our research to correlated areas. Specially, we advanced on music data science. In essence, music data is time series. Specifically, when assessing music data, the usual procedure is to extract features using sliding windows, resulting in multidimensional time series.

Comparing music data with similarity methods is a very common approach. For instance, we used distances between subsequences (music excerpts) as intermediate step for semi-supervised learn of music genres [Silva et al. 2014] and as the core of cover song identification systems [Silva et al. 2015b, Silva et al. 2016b].

From the methods proposed in this thesis, we highlight the Similarity Matrix Profile (SiMPle), an adaptation of the Matrix Profile to music applications [Silva et al. 2016b]. As well as the MP complies with many time series mining tasks, we presented simple algorithms that use SiMPle to accomplish several different music mining tasks.

5. Summary of Contributions

The work performed during this Ph.D. research resulted in a variety of contributions. As consequence, it resulted in a diversity of papers published in high-impact conferences and journals. In this section, we briefly present these publications.

The conferences selected to report our advances are among the most prominent venues on both data mining and music data science domains. The results on improving DTW for time series mining were presented in the SIAM SDM [Silva and Batista 2016] and IEEE ICDM [Silva et al. 2016a]. Also, there is a paper under review for the IJCAI[1]. The results on music data science were all published in the ISMIR Conference [Silva et al. 2014, Silva et al. 2015b, Silva et al. 2016b]. All these conferences are indexed as A1 by the Qualis-CAPES.

Moreover, some of the contributions of our work were submitted to relevant journals. The advances on similarity search are presented in an accepted manuscript to the DAMI journal (impact factor: 3.160) [Silva et al. ress]. The fast algorithm to calculate SiMPle and some other extensions of it are in a manuscript under review in the IEEE TMM (impact factor: 3.509)[2]. Additionally, the wide and deep review on DTW and its use in time series mining made for the thesis is being compiled as a survey, submitted to the ACM Computing Surveys. These journals are indexed as A1 in the Qualis-CAPES. Besides, in an early stage of this research, we mixed feature extraction and similarity-based

[1]Silva, D. F. and Batista, G. E. A. P. A. (under review). Elastic Time Series Motifs and Discords. In *International Joint Conference on Artifical Intelligence*.

[2]Silva, D. F., Yeh, C.-C. M., Zhu, Y., Batista, G. E. A. P. A., and Keogh, E. (under review). Fast Similarity Matrix Profile for Music Analysis and Exploration. *IEEE Transactions on Multimedia*.

methods to classify insects by species from digital signals [Silva et al. 2015a], which was published by the JINT (impact factor: 1.512, Qualis B1).

In addition to these publications, additional work was done in contributions, mostly on topics intrinsically related to this thesis. One example is the aforementioned Matrix Profile. This work resulted in two publications, namely one in conference [Yeh et al. 2016] and the other in journal [Yeh et al. 2017], both indexed as A1. Moreover, the candidate collaborated in efforts for time series classification [Souza et al. 2014, Giusti et al. 2015, Giusti et al. 2016, Dau et al. 2017] (conferences indexed as A2, not indexed – h-index: 12 –, B1, and not indexed – h-index: 25–, respectively). The latter has an extension submitted to the DAMI journal (Qualis A1)[3].

Finally, the candidate also collaborated on domains less related to the main topic of the thesis. Specially, the candidate has publications on anytime algorithms [Lemes et al. 2014] (conference, Qualis B1) and data stream classification in scenarios where the true label is not available in the deployment phase [Souza et al. 2015a, Souza et al. 2015b] (conferences Qualis B1 and A1, respectively).

6. Concluding Remarks

This thesis achieved several significant results in time series mining by similarity and music data science. In both cases, we introduced algorithms that improved the state-of-the-art in both efficiency and efficacy. Besides, all the methods proposed in this thesis are simple to implement, as well as space and time-efficient. More important, we made available all the source codes and data used to develop and evaluate our methods, so our advances could be reproducible and further applied by researchers and practitioners.

As future work, we intend to extend the improvements on the DTW algorithm to other kinds of data, like spatio-temporal data. Also, we are extending the main idea of pruning the DTW's algorithm to other dynamic programming-based methods, which may have impact in domains such as optimization and string search. Lastly, we intend to intensify the research on music data science, initially based on the method SiMPle.

References

Dau, H. A., Silva, D. F., Petitjean, F., Forestier, G., Bagnall, A., and Keogh, E. (2017). Judicious setting of dynamic time warping's window width allows more accurate classification of time series. In *IEEE BigData Conference*.

Giusti, R., Silva, D. F., and Batista, G. E. A. P. A. (2015). Time series classification with representation ensembles. *Lecture Notes in Computer Science*, 9385:108–119.

Giusti, R., Silva, D. F., and Batista, G. E. A. P. A. (2016). Improved time series classification with representation diversity and svm. In *IEEE International Conference on Machine Learning and Applications*, pages 1–6.

Lemes, C. I., Silva, D. F., and Batista, G. E. (2014). Adding diversity to rank examples in anytime nearest neighbor classification. In *IEEE International Conference on Machine Learning and Application*, pages 129–134.

[3]Dau, H. A., Silva, D. F., Petitjean, F., Forestier, G., Bagnall, A., and Keogh, E. (under review). Optimizing Dynamic Time Warping's Window Width for Time Series Data Mining Applications. *Data Mining and Knowledge Discovery*.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 262–270.

Silva, D. F. and Batista, G. E. A. P. A. (2016). Speeding up all-pairwise dynamic time warping matrix calculation. In *SIAM International Conference on Data Mining*, pages 837–845.

Silva, D. F., Batista, G. E. A. P. A., and Keogh, E. (2016a). Prefix and suffix invariant dynamic time warping. In *IEEE International Conference on Data Mining*, pages 1209–1214.

Silva, D. F., Giusti, R., Keogh, E., and Batista, G. E. A. P. A. (in press). Speeding up similarity search under dynamic time warping by pruning unpromising alignments. *Data Mining and Knowledge Discovery*, pages 1–32.

Silva, D. F., Rossi, R. G., Rezende, S. O., and Batista, G. E. A. P. A. (2014). Music classification by transductive learning using bipartite heterogeneous networks. In *International Society for Music Information Retrieval Conference*, pages 113–118.

Silva, D. F., Souza, V. M., Ellis, D. P., Keogh, E. J., and Batista, G. E. (2015a). Exploring low cost laser sensors to identify flying insect species. *Journal of Intelligent & Robotic Systems*, 80(1):313–330.

Silva, D. F., Souza, V. M. A., and Batista, G. E. A. P. A. (2015b). Music shapelets for fast cover song regognition. In *International Society for Music Information Retrieval Conference*, pages 441–447.

Silva, D. F., Yeh, C.-C. M., Batista, G. E. A. P. A., and Keogh, E. (2016b). SiMPle: assessing music similarity using subsequences joins. In *International Society for Music Information Retrieval Conference*, pages 23–29.

Souza, V. M. A., Silva, D. F., and Batista, G. E. A. P. A. (2014). Extracting texture features for time series classification. In *International Conference on Pattern Recognition*, pages 1425–1430.

Souza, V. M. A., Silva, D. F., Batista, G. E. A. P. A., and Gama, J. (2015a). Classification of evolving data streams with infinitely delayed labels. In *IEEE International Conference on Machine Learning and Applications*, pages 214–219.

Souza, V. M. A., Silva, D. F., Gama, J., and Batista, G. E. A. P. A. (2015b). Data stream classification guided by clustering on nonstationary environments and extreme verification latency. In *SIAM International Conference on Data Mining*, pages 873–881.

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., and Keogh, E. (2016). Matrix Profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *IEEE International Conference on Data Mining*, pages 1317–1322.

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Zimmerman, Z., Silva, D. F., Mueen, A., and Keogh, E. (2017). Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, pages 1–41.