Abstract

Time series data can be pervasively found in many fields. We are always interested in retrieving knowledge from existing data. Finding similar subsequences could be the first step before preforming any analyses. Dynamic time warping is one of the best distance functions to measure similarity between two subsequences. However, the time complexity for computation of dynamic time warping is $O(N^2)$ for a query of length N, which is slow to calculate. To accelerate the computation, one method is to use lower bound functions. *LB_Keogh* and its variants are the most popular lower bound functions, which have time complexity of $O(N)$.

In this dissertation, we will focus on exact sequential search on normalized time series sequences under dynamic time warping for large dataset, with the assumption that dataset is non-segmented and lower bound function *LB_Keogh* is used. The state-of-art method is UCR Suite.

The contribution of this dissertation is to improve the efficiency of UCR Suite under the scenario that (i) we are interested in finding similar subsequences for a few arbitrary-length queries, but not a single query, and (ii) the lengths of queries are long. A new lower bound function, namely *LB_LowResED*, is introduced. It is a lower bound function of *LB_Keogh*. The two assumptions are usually true for financial data analysis. Users would like to ask a few arbitrary-length long queries on the same set of financial data.

This dissertation is composed of four parts: (i) introduction of time series sequences searching, (ii) related works, especially for the state-of-art method UCR Suite, (iii) introduction of the new lower bound function *LB_LowResED*, and (iv) experiment results.

Table of Contents

Statement of Authorship .... **I**

Acknowledgements .... **II**

Abstract .... **III**

Table of Contents .... **IV**

List of Figures .... **VI**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Problem Definition 5

Chapter 2 Related Works **7**

2.1 Sequential Search for Arbitrary-Length Queries 7

2.2 Indexing Search for Arbitrary-Length Queries 10

2.2.1 Tree Index 10

2.2.2 Key-value Index 12

2.3 State-of-Art Method: UCR Suite 13

2.3.1 Cascading lower bounds 13

2.3.2 Online z-normalization 14

2.3.3 Reordering of Comparison for Early Abandon 15

Chapter 3 New Method: LB_LowResED **17**

3.1 *LB_PAA*: Lower bound for fixed-length queries 17

3.2 *LB_LowResED*: Lower bound for arbitrary-length queries 19

3.2.1 Renormalization 20

3.2.2 Shifting Windows 28

3.3 Time complexity analysis 30

3.4 Limitations of *LB_LowResED* 32

Chapter 4 Experiments **33**

4.1 FX Market Data 33

4.2 Power Consumption Data 36

4.3 Changing block length for FX Market Data 37

Chapter 5 Conclusion **38**

Reference **39**

List of Figures

Figure 1: Example of Euclidean distance and dynamic time warping 2

Figure 2: Example of normalizing subsequences for different length queries. 4

Figure 3: Example of dynamic time warping, *LB_Kim* and *LB_Keogh* 8

Figure 4: Early abandon of subsequence under Euclidean distance.....................................9

Figure 5: Piecewise Aggregate Approximation as lower bound of Euclidean distance..... 11

Figure 6: Example of reordering in UCR Suite... 15

Figure 7: Example of *LB_Keogh* and *LB_PAA* of *LB_Keogh*.. 17

Figure 8: Example of low-resolution sequence 20

Figure 9: Example of upper/ lower bound of truncated *MBR* 21

Figure 10: Example of S^U and S^L 22

Figure 11: Illustration of minimum possible μ_S...24

Figure 12: Example of *MBR* distance...27

Figure 13: Example of low-resolution query and subsequence..29

Figure 14: Example of shifting window and its solution..29

Figure 15: Query in Test #1 for FX Market Data...33

Figure 16: Results of Test #1 - Searching Time for FX Market Data 34

Figure 17: Results of Test #1 - Pruning Ratio & Time Saved with *LB_LowResED* 34

Figure 18: Results of Test #2 - Searching Time for FX Market Data 35

Figure 19: Results of Test #2 - Pruning Ratio & Time Saved with *LB_LowResED* 35

Figure 20: Results of Power Consumption Data - Searching Time......................................36

Figure 21: Results of Power Consumption Data - Pruning Ratio & Time Saved with *LB_LowResED*..36

Figure 22: Results of FX Market Data Test #1 – Various block lengths 37

Chapter 5 Conclusion

The state-of-art solution for similarity search for time series subsequence under dynamic time warping is UCR Suite. UCR Suite suggests that lower bounds of dynamic time warping could be *LB_Keogh* and *LB_Kim*, which are measured in Euclidean distance. These cascading lower bounds are tested before dynamic time warping is computed. UCR Suite introduces two techniques to improve the efficiency of Euclidean distance calculation, namely online z-normalization and reordering.

The new lower bound function, *LB_LowResED*, is designed for accelerating sequential search of time series data if similarity is measured in Euclidean distance. It is a lower bound for Euclidean distance, which can be used as a lower bound of *LB_Keogh*. In this dissertation, we modify UCR Suite by inserting this new cascading lower bound *LB_LowResED*. It uses low-resolution technique to improve the speed of Euclidean distance computation. By experiments, we show that *LB_LowResED* could improve Euclidean distance computation for long queries.

In this dissertation, we improved UCR Suite by modifying the cascading lower bound technique. For future works, we could attempt to improve UCR Suite by modifying early abandon and reordering techniques. One direction is to study the relation between shape of query, block length of *LB_LowResED* and the effectiveness of acceleration.

Note that both naïve UCR Suite and *LB_LowResED* use the early abandon and reordering to improve the searching speed. Usefulness of the early abandon technique and reordering technique are highly depending on the shape of query. It is possible to further improve UCR Suite by studying the optimal early abandon and reordering strategies.

Reference

[1] Berndt, Donald J., and James Clifford. "Using dynamic time warping to find patterns in time series." KDD workshop. Vol. 10. No. 16. 1994.

[2] Tselas, Nikolaos, and Panagiotis Papapetrou. "Benchmarking dynamic time warping on nearest neighbor classification of electrocardiograms." Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments. ACM, 2014.

[3] Sakoe, Hiroaki, and Seibi Chiba. "Dynamic programming algorithm optimization for spoken word recognition." IEEE transactions on acoustics, speech, and signal processing 26.1 (1978).

[4] Rath, Toni M., and Raghavan Manmatha. "Word image matching using dynamic time warping." 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. Vol. 2. IEEE, 2003.

[5] Keogh, Eamonn, et al. "Indexing large human-motion databases." Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004.

[6] Mager, Johannes, Ulrich Paasche, and Bernhard Sick. "Forecasting financial time series with support vector machines based on dynamic kernels." 2008 IEEE Conference on Soft Computing in Industrial Applications. IEEE, 2008.

[7] Rakthanmanon, Thanawin, et al. "Searching and mining trillions of time series subsequences under dynamic time warping." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[8] Echihabi, Karima, et al. "The lernaean hydra of data series similarity search: An experimental evaluation of the state of the art." Proceedings of the VLDB Endowment 12.2, 2018.

[9] Keogh, Eamonn, and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." Knowledge and information systems 7.3 (2005).

[10] Li, Yuhong, Man Lung Yiu, and Zhiguo Gong. "Discovering longest-lasting correlation in sequence databases." Proceedings of the VLDB Endowment 6.14 (2013).

[11] Linardi, Michele, and Themis Palpanas. "Scalable, variable-length similarity search in data series: The ULISSE approach." Proceedings of the VLDB Endowment 11.13 (2018).

[12] Kim, Sang-Wook, Sanghyun Park, and Wesley W. Chu. "An index-based approach for similarity search supporting time warping in large sequence databases." Proceedings 17th International Conference on Data Engineering. IEEE, 2001.

[13] Wu, Jiaye, et al. " KV-Match: A Subsequence Matching Approach Supporting Normalization and Time Warping." ICDE 2019: 866-877.

[14] Keogh, Eamonn, et al. "Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures." *The VLDB journal* 18.3 (2009): 611-630.

[15] Lemire, Daniel. "Faster retrieval with a two-pass dynamic-time-warping lower bound." *Pattern recognition* 42.9 (2009): 2169-2180.