Time Series Clustering Based on Frequency Distribution of First-order Differences

P. Thanakulkairid, T. Trakulthongchai, P. Vatiwutipong*

Introduction

The recent rise of big data analytics in policy decision, business strategy, and biomedicine generates the need for swift and accurate algorithms for data analysis, such as prediction, classification, and evaluation. One of the most prominent types of data in everyday life is time series, a type of data that is, by definition, collected over time in consecutive, equal intervals.

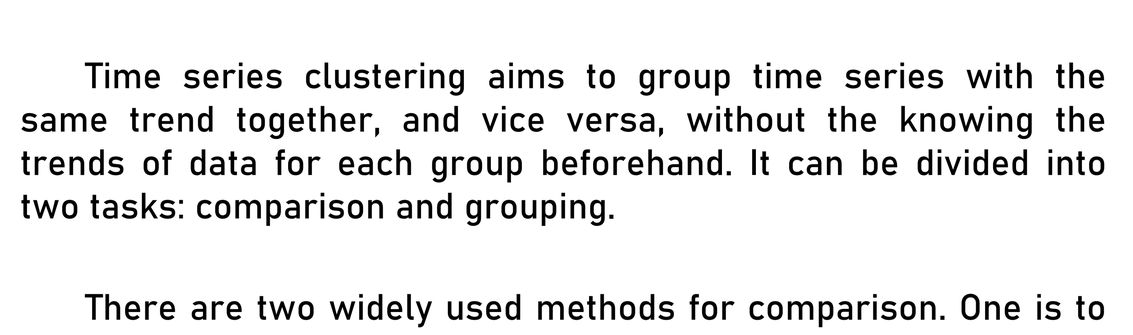
Time series clustering aims to group time series with the same trend together, and vice versa, without the knowing the trends of data for each group beforehand. It can be divided into two tasks: comparison and grouping.

There are two widely used methods for comparison. One is to use statistical distances to compare time series directly, which has the advantage of speed from its linear time complexity but is relatively inaccurate. Another is dynamic time warping (DTW) algorithm, which brings the desired accuracy at the expense of higher (quadratic) time complexity, causing it to be inapplicable for large data set.

Objectives

- Find fast and accurate algorithm for time series clustering using statistical distance to compare the frequency distribution of first-order differences
- Determine the optimal statistical distance for time series clustering using frequency distribution of first-order differences

Methodology



Euclidean Distance

$$ED(P,Q) = \sqrt{\sum_{x\in X}\big(P(x)-Q(x)\big)^2}$$

Jensen divergence

$$JD(P,Q) = \sum_{x\in X}\big(P(x)-Q(x)\big)\log\frac{P(x)}{Q(x)}$$

Wasserstein distance

$$WD(P,Q) = \sum_{x\in X}|P'(x)-Q'(x)|$$

Bhattacharyya distance

$$BD(P,Q) = -\log\sum_{x\in X}\sqrt{P(x)Q(x)}$$

Algorithm 1 Clustering.
This algorithm clusters N time series into m clusters.

Step 1: For each time series A_i with length L_i, find the first-order difference sequence D_i where each entry is the difference of two consecutive entries of A_i.

Step 2: Let $D = \cup D_i$. Define a probability space for first-order differences X where each element i corresponds to a range of values R_i of length equal to half the standard deviation of D such that R_1 corresponds to the range with lowest values and increases until all possible values of D are accounted for. Assign to each time series a matching probability mass function f_i such that

$$f_i(k) = C_k/(L_i - 1)$$

where C_k counts the number of elements in D_i.

Step 3: For each couple i and j, define the difference between A_i and A_j to be $DF(A_i, A_j)$ where $DF = ED, JD, WD, BD$.

Step 4: Use these distance values in partition around medoids (PAM) algorithm with m medoids as described by [3].

Results & Discussion

The result of the first part of the first experiment is shown in Figure 1. We can observe the JD yields the most accuracy, although it is comparable to ED and BD for $p \geq 90$ and $p \leq 50$, and all the mentioned distances perform better than the DTW controls. For $50 < p < 90$ JD is significantly better than the others, which illustrates the higher sensitivity of its similarity detection.

Table 1 shows the results of the last part of Experiment 1 and Experiment 2. For both experiments, JD has the most accuracy on average of all distance functions and controls. However, ED and WD have better best-case performance in Experiment 1.

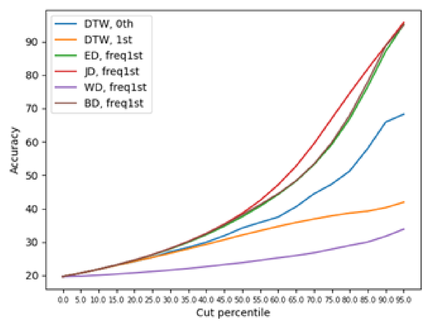


Figure 1 – Accuracy from each distance function and controls for each tested cut percentile p.

Distance functions	Exp. 1 AA (%)	Exp. 1 OA (%)	Exp 2. AA (%)	Exp 2. OA (%)
ED	73.76	89.60	43.09	50.57
JD	79.78	87.00	72.61	76.24
WD	36.54	40.20	N/A	N/A
BD	75.26	93.20	42.27	48.29
DTW_0	61.90	70.64	51.29	58.53
DTW_1	50.65	58.40	56.79	64.12

Table 1 – The results of Experiment 1 last part and Experiment 2. AA and OA are average accuracy and optimal accuracy, respectively.

Conclusion

In this research we have constructed an algorithm for time series clustering based on first-order differences of the time series, that is Algorithm 1. In addition, we have determined that the optimal distance function for comparing the PMFs of the first-order differences in Jensen's divergence.

Future Work

There are multiple avenues for extension of this research.

- extend the study from univariate time series to multivariate time series
- extend the study from first-order differences to second-order or higher order differences
- explore with extrapolation of time series data and clustering using the first derivative of the extrapolated curve

References

[1] Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information systems, 53* 16-38.

[2] Wu, R., & Keogh, E. J. (2022). FastDTW is Approximate and Generally Slower Than the Algorithm it Approximates. *IEEE Transactions on Knowledge and Data Engineering, (34)* 8, 3779-3785.

[3] Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications, 36*(2), 3336-3341.

[4] Zhang, X., Wu, J., Yang, X., Ou, H., & Lv, T. (2009). A novel pattern extraction method for time series classification. *Optimization and Engineering, 10*(2), 253-271.