Writing Project Summaries

Eliot Liucci

2023-09-20

Article Name: Hydrological Time Series Clustering: A Case Study of Telemetry Stations in Thailand (Prakaisak and Wongchaisuwat 2022)

Article Summary

This article is based on time series data that comes from over 80 telemetry stations in Thailand that are used to monitor water levels. The author's goal is to cluster the telemetry stations using various clustering algorithms and dissimilarity measures. This relates to my topic for my writing project as I am interested in doing a survey of various statistical learning algorithms. The motivation for this study was to dive deep into the clustering process as clustering is mainly a pre-analysis step that allows the simplification of the data into a few key groups. Instead of going through a single algorithm like some other paper's I have read, this one is fairly broad and covers the entire process of data cleaning, smoothing, feature representation, dimension reduction, and clustering for several different methods and then compares them all using a single quantitative summary called the Fowlkes-Mallows score, which represents the overall similarity within groups. The main takeaway from this article is that based on what type of data is being collected, there are different combinations of the previously mentioned steps that give the best results. For hydrological time series, they ended up discovering that using a Hierarchical Agglomerative Clustering (HAC) algorithm with a Euclidean distance dissimilarity measure is most effective.

Article Name: Time Series Clustering - A Decade Review (Aghabozorgi, Shirkhorshidi, and Wah 2015)

Article Summary

This paper gives a vast overview of various time series clustering techniques. They start by breaking down the 3 major techniques for time series clustering, which are shape-based, feature-based, and model-based. Each method can be optimal depending on the nature of the data, but based on other articles read, it seems shape-based methods are ideal for continuous time series. Feature-based methods involve extracting features from the time series (daily minumum, daily maximum, daily range, etc.) and clustering on those features rather than the raw measurements. Model-based methods involve fitting some formulation of an ARIMA model and then clustering based on features from that model (autocorrelation, seasonality, etc.). Shape-based, which is the most common method, uses the raw data itself to compare time series. This could be as simple as taking the difference of the measurements on a day between two time series. The author dives into distance metrics as well, the most common and recommended of which are Dynamic Time Warping, Pierson's Correlation Coefficient, and Euclidean Distance. The algorithms that the author recommends using are K-means partitioning and Hierarchical clustering.

Article Name: Clustering of time series data—a survey (Warren Liao 2005)

Article Summary

There are several popular methods to time series clustering that are similar to static data clustering. Most of those methods are covered in previous papers, but a few stand out. This author goes more into detail about the major types of Hierarchical Modeling: Agglomerative and Divisive. Agglomerative starts with each unit (in my case, a time series), and then groups the nearest 2 together, then the next nearest comparison, and so on. This continues until there is only 1 cluster. Divisive would do the same but in reverse order, starting with 1 cluster and continuing until there are n, assuming there are n time series being compared. These methods can differ vastly based on the linkages used. Linkages are covered well in another textbook I have cited (Introduction to Statistical Learning).

Week 4 Assignment Questions

Of the three papers you have read, which has been the most relevant for the writing project topic (or topics) you are interested in pursuing? Briefly explain why.

The paper that gave me the most information on the topic is "Time Series Clustering - A Decade Review" (Aghabozorgi, Shirkhorshidi, and Wah 2015). This paper covers every component that I was interested in looking at, so it gives me a really strong starting point as to where I really want to focus my attention for my project.

Are you still interested in pursuing the topic (or topics) you set out to pursue after reading these three papers? Briefly explain.

I am still interested in this topic, although rather than focusing on a broader overview of Time Series Clustering, I will likely focus on some popular combinations of distance metrics and clustering algorithms. This will allow me to go much more in depth into each algorithm with my exploration.

Have the papers peaked your interest about any other topics in statistics you would like to learn about? Briefly explain.

One of the papers got me thinking about the potential for Neural Networks. They mention that Neural Networks have become more popular as data is more readily available for an accurate training of the network.

References

Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. "Time-Series Clustering–a Decade Review." *Information Systems* 53: 16–38.

Everitt, Brian, and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with r.* Springer Science & Business Media.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An Introduction to Statistical Learning.* Vol. 112. Springer.

Maharaj, Elizabeth Ann, Pierpaolo D'Urso, and Jorge Caiado. 2019. *Time Series Clustering and Classification.* CRC Press.

Milligan, Glenn W, and Martha C Cooper. 1985. "An Examination of Procedures for Determining the Number of Clusters in a Data Set." *Psychometrika* 50: 159–79.

Prakaisak, Intouch, and Papis Wongchaisuwat. 2022. "Hydrological Time Series Clustering: A Case Study of Telemetry Stations in Thailand." *Water* 14 (13): 2095.

Rani, Sangeeta, and Geeta Sikka. 2012. "Recent Techniques of Clustering of Time Series Data: A Survey." *International Journal of Computer Applications* 52 (15).

Warren Liao, T. 2005. "Clustering of Time Series Data—a Survey." *Pattern Recognition* 38 (11): 1857–74. https://doi.org/https://doi.org/10.1016/j.patcog.2005.01.025.