Identifying Time Series Similarity in Large-Scale Earth System Datasets

Payton A. Linton*
William M. Melodia*
Alina Lazar
palinton,wmmelodia@student.ysu.edu
alazar@ysu.edu
Youngstown State University
Youngstown, Ohio

Deborah Agarwal, Ludovico Bianchi
Devarshi Ghoshal, Kesheng Wu
Gilberto Pastorello, Lavanya Ramakrishnan
daagarwal,lbianchi,dghoshal,
kwu,gzpastorello,lramakrishnan@lbl.gov
Lawrence Berkeley National Laboratory
Berkeley, California

ABSTRACT

Scientific data volumes are growing every day and instrument configurations, quality control and software updates result in changes to the data. This study focuses on developing algorithms that detect changes in time series datasets in the context of the Deduce project. We propose a combination of methods that include dimensionality reduction and clustering to evaluate similarity measuring algorithms. This methodology can be used to discover existing patterns and correlations within a dataset. The current results indicate that the Euclidean Distance metric provides the best results in terms of internal cluster validity measures for multi-variable analyses of large-scale earth system datasets. The poster will include details on our methodology, results and future work.

ACM Reference Format:
Payton A. Linton, William M. Melodia, Alina Lazar, Deborah Agarwal, Ludovico Bianchi, Devarshi Ghoshal, Kesheng Wu, and Gilberto Pastorello, Lavanya Ramakrishnan. 2019. Identifying Time Series Similarity in Large-Scale Earth System Datasets. In *SC '19: The International Conference for High Performance Computing, Networking, Storage, and Analysis, November 19–22, 2019, Denver, CO.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Scientific datasets are continuously increasing in size and diversity and it is becoming imperative to develop methods to quantitatively assess and track changes in such datasets. In many cases, these large datasets have spatial and/or time components and may vary in terms of resolution, quality, and availability. It is important for researchers to study the changes, variations, and patterns that occur between different versions of data and within the same version. A more in depth understanding of the data leads not only to better analysis and research results, but can save the need to re-analyze the datasets when changes are insignificant. The main goal of this research is to develop statistical, machine learning and visualization methods to enable cross-comparison of large scientific datasets that have spatial and temporal components in the context of the Deduce project[1]. In this context, we focus on the deployment and evaluation of combinations of similarity measures, non-linear dimension reduction and clustering algorithms applied to high-dimensional time series datasets to help with similarity searches.

2 RELATED WORK

Previous work on similarity measures for time series data have focused on single variable and multivariate time series classification [7] but not clustering. Similarity measures computed based on sequence alignment methods provide better results compared to the standard Euclidean distance. Dynamic time warping (DTW) [2] is an algorithm for measuring similarity between two temporal sequences, using pairwise sequence alignment which performs better than other similarities in terms of classification accuracy. However, this algorithm is based on dynamic programming and takes n^2 computations. The computed similarity matrix, is used as input into the clustering procedure [1] to group and find representative patterns. The quality of the clustering can be evaluated using internal clustering validity measures [3]. In addition, dimensionality reduction methods such as UMAP [4] are good tools not only for visualizing the clustering results but also to embed the time series data [5, 8] in a low dimensional space to improve the results.

3 METHODS

Clustering together with visualization methods can help scientists understand the underlining structure existing in large unlabeled datasets. Clustering is an unsupervised machine learning algorithm that groups together similar time series while maximizing the distance between groups. A crucial step in clustering approaches dealing with time series is to decide how to define similarity. This project evaluates two issues facing unlabeled time series clustering analysis: the choice of similarity measure and the effects of using dimensionality reduction embeddings. Changes in clustering solutions are systematically assessed in a full combination of experiments designed so that the two types of problems can be examined in relation to each other since they affect the clustering solution.

All the experiments were applied to the FluxNet scientific dataset. It includes data collected at sites from multiple regional flux networks. this extensive global network of towers (over 800) provides the largest produced dataset of CO2, water vapor, energy fluxes

[*] Both authors contributed equally to this research.

[1] http://deduce.lbl.gov

SC '19, November 19–22, 2019, Denver, CO
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

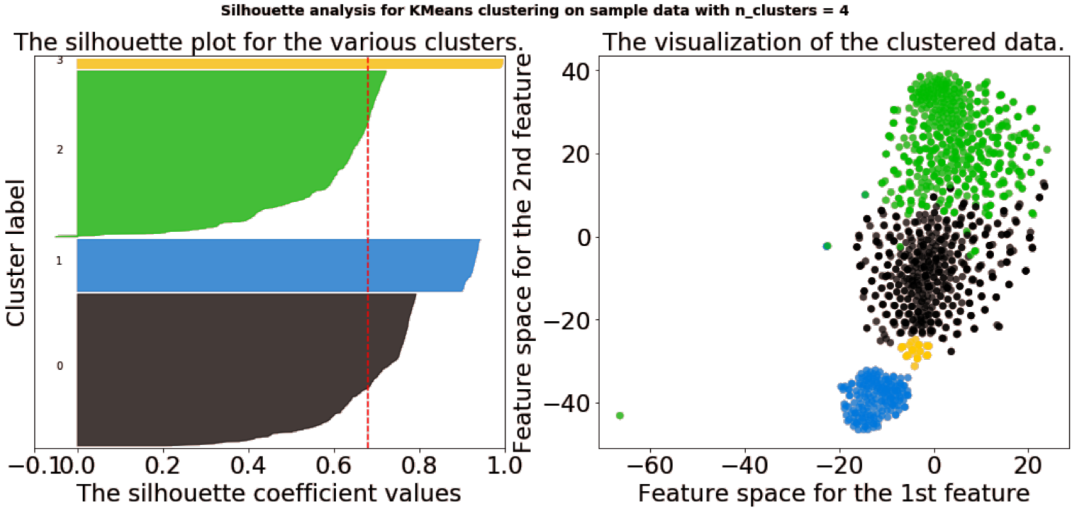

Figure 1: Silhouette Analysis for K-Means Clustering and UMAP Visualization with Cluster Labels

and net ecosystem exchange (NEE). Bio-geoscientists use this data to understand the essential factors that influence the CO2 and NEE fluxes that critical for our understanding of relevant weather and climate patterns. We validate our proposed methods on data extracted from the FluxNet2015 dataset [6]. We apply our methods to both climate and ecosystem variables and combinations of these.

We address the problem of non-existing classification labels by evaluating the methodology in terms of several internal clustering measures and by examining the distribution of sites plus years combinations in the proposed solutions. The nonlinear dimensionality reduction technique, UMAP was applied to visualize the solutions and also to provide data embeddings in a lower dimensionality space for better clustering solutions. These embeddings are easier to cluster in comparison with the pre-computed similarity matrices.

4 EARLY RESULTS

The results we obtained in terms of the three cluster validity measures are very similar, we found that the standard Euclidean distance works best when clustering is applied directly to the pre-computed similarity matrix, while DTW provides better results when applied to UMAP embeddings. First, we used the following variables to run experiments: temperature, shortwave radiation, precipitation, wind speed, and vapor pressure deficit (VPD). After running K-Means clustering, ranging from 2 to 10 cluster solutions, the best number of clusters based on the validity measures is 4. It is interesting to note that the CHS measure keeps increasing as the number of cluster increases, while both ASW and DBS had almost constant values between 3 and 10 clusters, therefore any clustering solution could provide insights about the data patterns.

The Silhouette plot in Figure 1. proves the cluster solution is valid, since the silhouette coefficients of all the clusters are higher than the average value. Checking the cluster distribution in terms of the latitude and longitude of the site and the year the data was recorded, we know specific information about every cluster. The yellow cluster contains only sites near the US-Mexico border, at about 30°N. The grey cluster mainly consists of sites between 35°and 50°N. The green cluster consists of sites mainly above 50 °N, but also include some sites at high elevations. The main difference between the green and grey cluster is that the peak temperature is lower for the sites in the green cluster. The blue cluster contains sites mainly between 0°and 35°S, with the majority of the sites being in Australia. For temperature, shortwave radiation, and VPD the yellow cluster has the highest peak overall and always has higher values than the green and grey clusters, while the blue cluster follows the opposite patterns as the others due to the seasons being opposite in the southern hemisphere. For precipitation, we see the the yellow and blue clusters get precipitation in waves while the grey and green have consistent precipitation year round. For wind speed, all of the clusters follow about the same trend except the blue cluster, which has a unique trend. These variables are dependant on the season and latitude primarily, and then on vegetation type and altitude.

5 CONCLUSIONS

After testing multiple similarity algorithms including the Euclidean distance, Dynamic-Time Warping, Pearson Correlation, and Fourier Coefficients, we concluded that Euclidean distance gave the best on multi-variable data. Thus, the Euclidean distance was chosen as the distance measure for this combination of variables. This resulted in better cluster solutions with more interpretable results. These methods combined will allow for a more complete picture of similarities between the different sites and different years.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce and Development for Teachers and Scientists (WDTS) under the Visiting Faculty Program (VFP) program and of Advanced Scientific Computing Research under Contract No. DE-AC02-05CH11231 We also used resources at the National Energy Research Scientific Computing Center.

REFERENCES

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering – A decade review. *Inf. Syst.* 53 (Oct. 2015), 16–38.

[2] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, Vol. 10. aaai.org, 359–370.

[3] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. 2010. Understanding of Internal Clustering Validation Measures. In *2010 IEEE International Conference on Data Mining*. ieeexplore.ieee.org, 911–916.

[4] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (Feb. 2018). arXiv:stat.ML/1802.03426

[5] Minh Nguyen, Sanjay Purushotham, Hien To, and Cyrus Shahabi. 2017. m-TSNE: A Framework for Visualizing High-Dimensional Multivariate Time Series. (Aug. 2017). arXiv:cs.LG/1708.07942

[6] Gilberto Pastorello, Dario Papale, Housen Chu, Carlo Trotta, Deborah Agarwal, Eleonora Canfora, Dennis Baldocchi, and Margaret Torn. 2017. A new data set monitors land-air exchanges. *EOS* 98, 8 (2017), 27–32.

[7] Joan Serrà and Josep Lluis Arcos. 2014. An Empirical Evaluation of Similarity Measures for Time Series Classification. (Jan. 2014). arXiv:cs.LG/1401.3973

[8] Pattreeya Tanisaro and Gunther Heidemann. 2019. Dimensionality Reduction for Visualization of Time Series and Trajectories. In *Image Analysis*. Springer International Publishing, 246–257.