Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Feature selection for Time Series Clustering: A case study on Dengue in Peru

Maria Giohanna Martinez,[1] Diego H. Stalder,[2] Juan Vicente Bogado,[3] Christian E. Schaerer[4]
National University of Asuncion, Paraguay
Max Ramírez-Soto M.,[5] Denisse Champin[6]
Technological University of Peru, Lima, Peru

In recent decades, the world has experienced a health crisis due to the increase of infectious diseases cases, such as COVID-19, Dengue, Zika, among others. Dengue is one of the world's most important neglected tropical disease transmitted by vectors, mainly *Aedes Aegypti*. However, its alarming geographical expansion and its high economic impact on health care systems has drawn attention of decision makers. Prevention efforts requires accurately identification of geographical similarities and heterogeneities in dengue incidence patterns. In Peru, dengue cases rises to 68.000 cases from 2017. We consider weekly reported cases, in 376 districts of Peru during 2020. The dataset is build from the *CDC's Health Situation* web portal [3].

Time series clustering algorithms can be used for detecting the geographical regions where the environmental conditions for mosquitoes combined with local social-economical activities may cause high dengue incidence. Clustering algorithm is organized in three approaches: shape-based, model-based and feature-based [1]. Time series clustering, specially feature-based ones improve short term forecasting in deep learning models [2]. There are several features used on time series [4], nevertheless, in case of diseases outbreaks (eg. dengue) is not clear which feature should be used to have meaningful clusters.

Elbow method is applied to define the number of clusters ($N = 7$ and fixed for all the experiments). To evaluate if [2] results hold on Peru data we perform the first experiment where three clustering algorithms and four similarity measures were applied to the raw time-series and features extracted from the data. The metrics considered in this experiment are the euclidean distance, correlation, spearman correlation and dynamic time warping. The performance evaluation of each algorithm and metric is calculated by the Silhouette score (similarly to [2]). In the second experiment, the feature selection, the data was represented by a six features vector were considered, e.g. the mean, variance, first order of auto-correlation, number of peaks, spectral entropy and number of crossing points (a sub-sample of the features considered in [4]). In order to identify the most important variables, we run the clustering algorithms 6 times, on each iteration one feature is removed. Finally results are ranked by the Silhouette score.

Table 1 presents the Silhouette score values where the columns indicates what feature based approach can give better clustering results. The rows indicates that for the Hierarchical clustering with the Euclidian distance and the Dynamic time warping have a better performance.

[1]giomartinezf@gmail.com
[2]dstalder@ing.una.py
[3]jvbogado@fctunca.edu.py
[4]cschaer@pol.una.py
[5]maxcrs22@gmail.com
[6]dchampin@utp.edu.pe

Table 1: Silhouette score values for each clustering algorithms and metric considered.

Algorithm	Metrics	Silhouette Score	
		Shape based	Feature based
Hierarchical	Euclidean distance	0.74553	**0.922099**
	Correlation	0.34810	0.6508807
	Spearman	0.34972	0.782121
	Dynamic time warping	0.73248	0.918868
K-means	Euclidean distance	0.74457	0.919916
	Correlation	0.36065	0.653263
	Spearman	0.40552	0.766690
	Dynamic time warping	0.65978	**0.924338**
DBScan	Euclidean distance	-0.23533	-0.372072
	Correlation	-0.13006	0.172792
	Spearman	0.21793	**0.862731**
	Dynamic time warping	-0.25206	-0.418219

Table 2: Top 2 combinations of selected features for each clustering algorithm.

Algorithm	Rank	Features selected	Silhouette score
Hierarchical	1	Mean, Var, ACF1, Peaks, Entropy	0.922270
	2	Var, ACF1, Peaks, Entropy, CPoints	0.922169
K-means	1	Mean, Var, ACF1, Peaks, Entropy	0.920582
	2	Var, ACF1, Peaks, Entropy, CPoints	0.920241
DBScan	1	Mean, Var, Peaks, Entropy, CPoints	0.955606
	2	Mean, Var, ACF1, Peaks, CPoints	0.951546

Table 2 shows the best two combinations of selected features for each clustering algorithm, evaluated by the Silhouette score. The best metric obtained in Table 1 is used. In all cases, when extracting the feature *CPoints* (following by extracting *Mean*) is obtained the best results. When DBScan is considered, if more features are considered, the more deteriorated is the silhouette score. This observation is in contrast with the other methods tested. Hence, at this moment, motivated by the results, the authors are working to improve the DBScan to be more competitive and extend the algorithms to consider more features related to the disease dynamics, such as transmission rate, recovery rate, susceptible population, among others.

Acknowledgement

We would like to show our gratitude to the Project funded by *Universidad Tecnológica de Perú* (P-2020-LIM-01) for their support during the course of this research. Christian E. Schaerer and Diego H. Stalder thanks FEEI-PROCIENCIA-CONACYT-PRONII.

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah. "Time-series clustering – A decade review". In: **Information Systems** 53 (2015), pp. 16–38. ISSN: 0306-4379. DOI: 10.1016/j.is.2015.04.007.

[2] J. V. Bogado et al. "Time Series Clustering to Improve Dengue Cases Forecasting with Deep Learning". In: **2021 XLVII Latin American Computing Conference (CLEI)** (2021), pp. 1–10. DOI: 10.1109/CLEI53233.2021.9640130.

[3] CDC. **Centers for Disease Control and Prevention Official Site**. Online. Accessed 10/02/2022, https://www.cdc.gov/.

[4] R. J. Hyndman, E. Wang, and N. Laptev. "Large-Scale Unusual Time Series Detection". In: **2015 IEEE International Conference on Data Mining Workshop (ICDMW)** (2015), pp. 1616–1619. DOI: 10.1109/ICDMW.2015.104.

