Romika Sairam

November 8, 2021

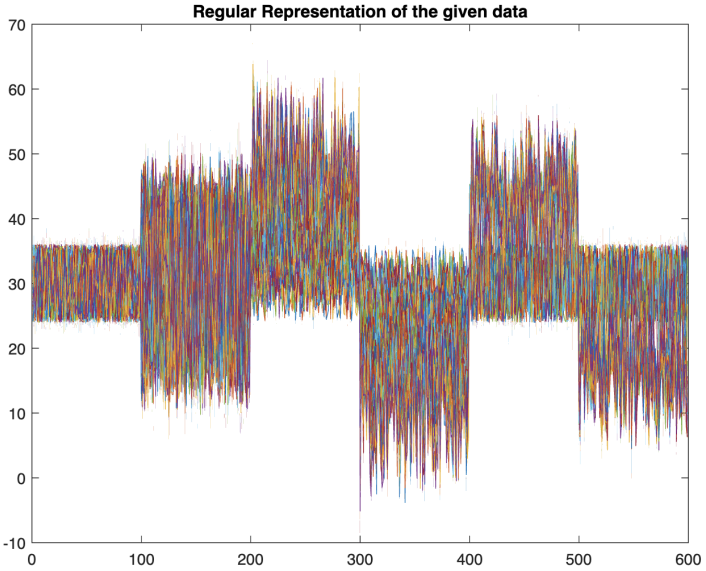
CSE454

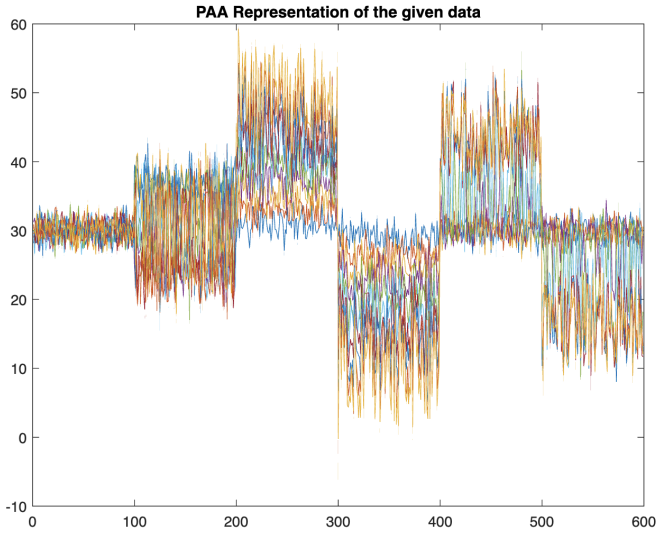
CSE454: Time Series Project
Final Report

The objective of the time series project is to take in a given dataset, read it in and create two data representations of the dataset using PAA and SAX. Here, the goal is to reduce the number of samples needed to represent the given data. The second piece of this project is the classification, which is done through Euclidean Distance and Manhatten Distance. Here, our goal is to make valid conclusions on which one is better and if there are any noticeable trends. The input in this case is the dataset that we are given and the outputs are the two sets which are the new representations of the datasets. The classifications of the dataset is also one of the expected outputs in this assignment.

Piecewise Aggregate Approximation

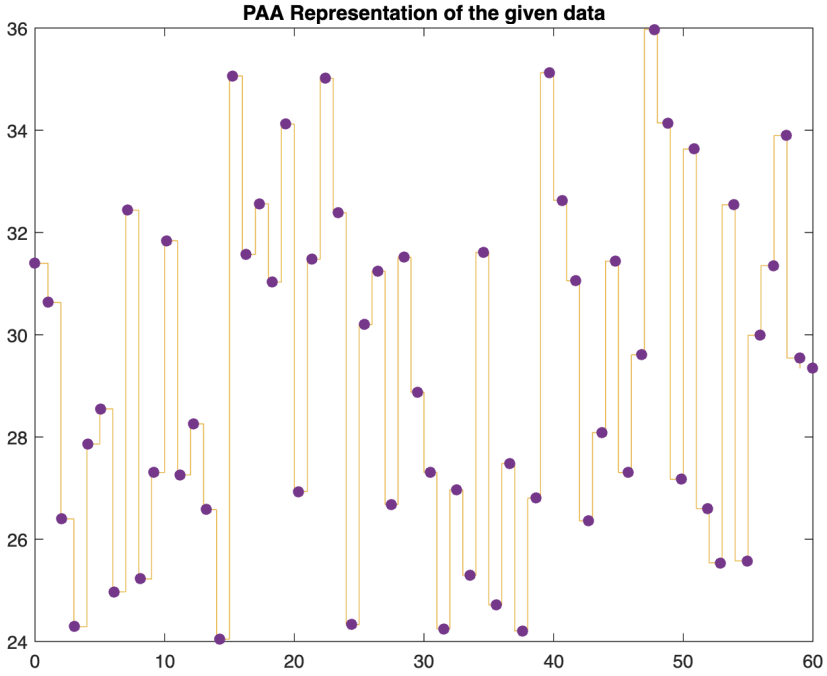
Piecewise Aggregate Approximation is a method whose aim is to reduce the data by approximating the aggregate values for each piecewise function within the bigger dataset. To implement this, the given dataset was divided into small windows to represent the piecewise functions. Then, each of the samples were assessed and if they fit within the range of the windows, the sample was stored in the PAA function. Finally, the aggregate value was taken of each of the samples. The graphs of the original dataset and the dataset after PAA was applied, is attached below:





From the images above, it is clearly displayed how PAA has reduced the number of samples as the original representation of the given data (image on left) is much more dense and bright than the PAA Representation (image on the left). This is a sample reduction.

For a more detailed scatterplot of the PAA Representation of the given data, the scatterplot is attached below:



Symbolic Aggregate Approximation

Symbolic Aggregate Approximation is another layer on top of the PAA method described above but in a SAX method, the samples are taken another set of aggregates on top of the result from the PAA method. This is to categorize and reduce the samples even more than the PAA method does to the original dataset.

Testing and Training datasets

Splitting the dataset into Training and Testing datasets is an important part of data analysis. To do this, the data is usually divided into unequal parts leaving the training dataset with a lesser portion of the data while the testing dataset has more of the dataset. This is to allow room for more testing than training for the model.

Classification

Classification is also another piece that is crucial to the given dataset. The given dataset, was split into six different classes namely, Normal, Cyclic, Increasing trend, Decreasing trend, Upward shift and Downward shift. This is implemented by splitting the given dataset into six parts where each of the parts contain 100 rows of the given data, which contains 600 rows of samples in total.

Conclusion

This project and the various changes made to the given dataset clearly displays how the methods PAA and SAX reduce the sample size required to represent the given data. SAX is a bit better than PAA due to its added layer of approximations on top of the PAA method.

References

Sairam, Romika. *CSE454 Time Series Project Code*.

https://github.com/romi-sai/Time_Series.git.

“Matlab.” *MATLAB Documentation*, https://www.mathworks.com/help/matlab/.