Time Series Project Report

Jared Murray

The overall objective of this project involved representing and classifying a given time series. The time series that was used was a synthetic control data set and two methods were used to represent it. One method was piecewise aggregate approximation (PAA) and the other using symbolic aggregate approximation (SAX). Both the raw data and the PAA data were then classified using Euclidean and Manhattan distance.

In generating the PAA representation of the synthetic control data, I first split each row of the data into 10 segments. Since there were 60 data points in each line/sample, each segment consisted of 6 points. The PAA representation of the sample was calculated by taking the average of each segment and therefore reducing the sample of 60 data-points to a sample of 10 data-points. Figure 1 below shows 6 different plots. Each plot shows a time series sample from a different class and both the original data and the PAA representation of the data are plotted. Figure 2 gives the descriptions of each class and corresponding rows within the original data set samples from each class are found.

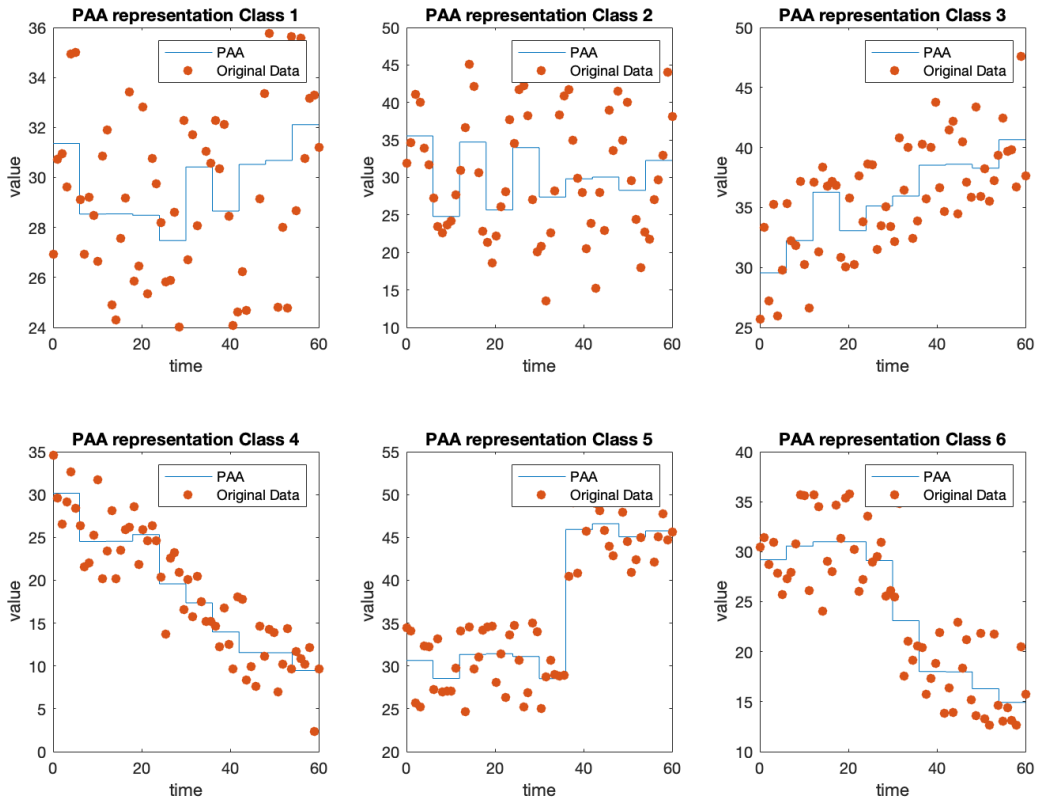

Figure 1

Class	Description	Rows
1	Normal	1 -100
2	Cyclic	101 – 200
3	Increasing trend	201 – 300
4	Decreased trend	301 – 400
5	Upward shift	401 – 500
6	Downward shift	501 – 600

Figure 2

The procedure followed in creating the SAX representation of the time series was to first normalize the data so that the standard deviation was 1 and mean 0. Then, the 10-segment PAA representation of normalized data was generated. The PAA values are then mapped to SAX symbols based on what range the values fall within. The symbols within the SAX alphabet used are the integers 1 through 10. Each sample/line of time series data is then represented as a string of 10 integers within the final SAX representation of the data. Figure 3 below shows 6 different plots. Similar to figure 2, each plot shows a time series sample from a different class and both the original data and the SAX representation of the data are plotted.

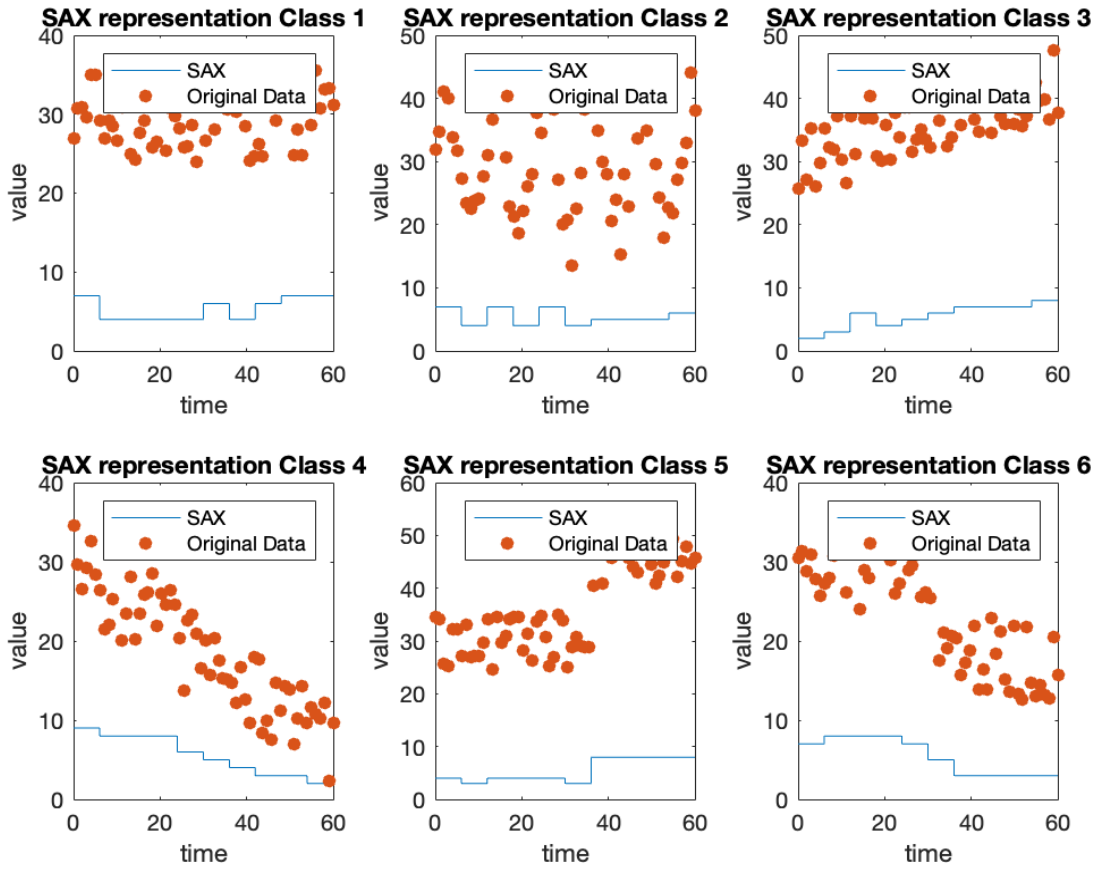

Figure 3

Both training and testing occurred on the full original or PAA data set. This was due to the way in which classification was done. Each sample/line of data within both the original set and its PAA representation was iterated through while the Euclidean and Manhattan distance were calculated between each given sample and all other samples. Each sample was classified using the same class as the sample that was the minimum distance away from it. Since two different distance measurements were used on both the original and PAA data sets, each sample was classified twice. Figure 4 therefore shows 4 different confusion matrices each generated with a different data representation/distance method combination. Each confusion matrix gives the number of times different samples from each class were classified using the given distance technique as its true class or a different class. Since there was only one more misclassification using Manhattan distance than with Euclidean distance for both the original and PAA sets, both distance methods were about equally effective on both sets. Classification using the PAA data representation, however, was more accurate than using the original set because more misclassifications were made with the original set. This indicates that reducing the size original data set using a method such as PAA may lead to better classification results.

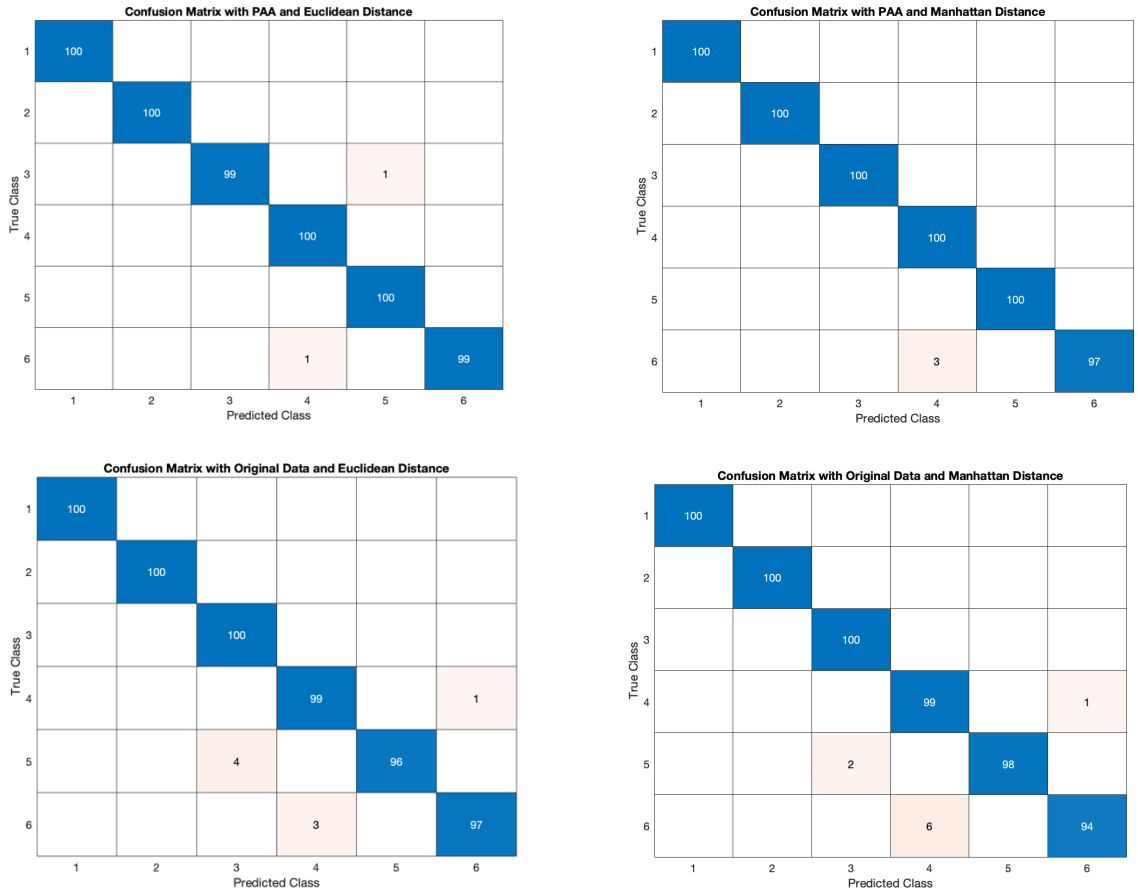

Figure 4

GIT LINK: https://github.com/Jared-01/TimeSeriesProject.git