CSE 454
Project 2

Time Series

Jesse Both

Fall 2021
November 7, 2021

Contents

Introduction

The purpose of this assignment was to explore representation and classification techniques of a time series. The data that was used for this assignment was a synthetic control data set. The representaion techniques that were used for this assigment were PAA - Piecewise Aggregate Approximation and SAX - Symbolic Aggregate Approximation. The classification was done by utilizing the Euclidean and Manhattan distance forumulas.

The Euclidean Distance Formula: $d(r,s) = \sqrt{\sum_{i=1}^{n}(r_i - s_i)^2}$
The Manhattan Distance Formula: $d(r,s) = \sum_{i=1}^{n}|r_i - s_i|$

Data

The provided data contains 6 classes. Each sample is contained within 1 row and has 60 columns. There are 100 samples of each type of class. The types of classes include:

- Normal
- Cyclic
- Increasing Trend
- Decreasing Trend
- Upward Shift
- Downward Shift

The class changes after 100 consecutive samples, so the first 100 samples are within the normal class followed by the next 100 sample being cyclic.

Preprocessing

Inorder to get the data into a more usable state, the entire data set was normalized using Matlabs normalize() function. This enables the data to be more uniform aswell as decreasing the distance between the upper and lower bounds.

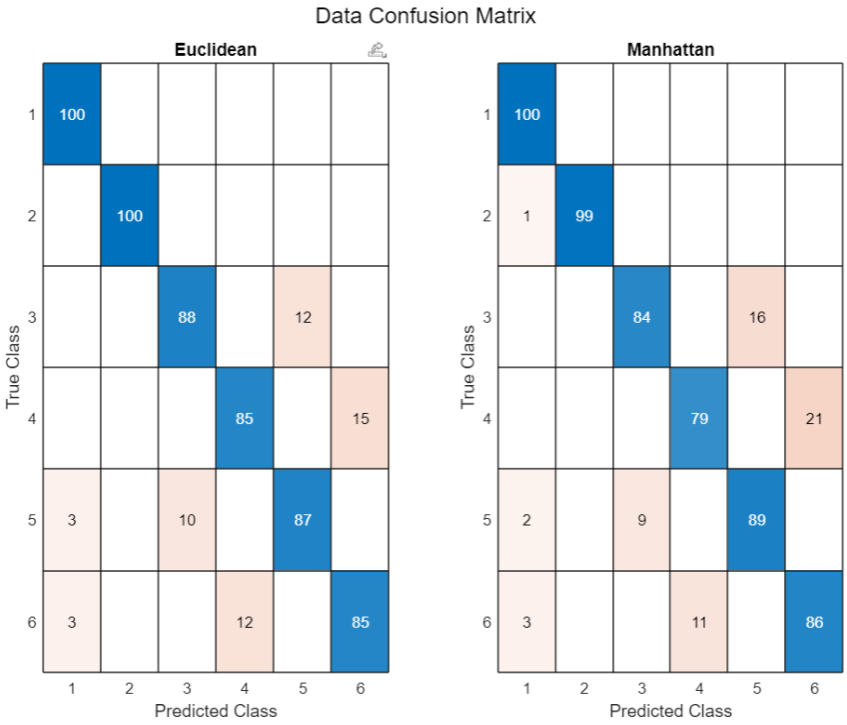
Training Data

To create a training data set, each class was seperated into their own vectors. Each column was then averaged to find the ideal sample based on the data. Once this was complete the training data would consist of a 6x60 matrix, 1 averaged row for each class.

Testing Data

The testing data that was used was the original, normalized data set.

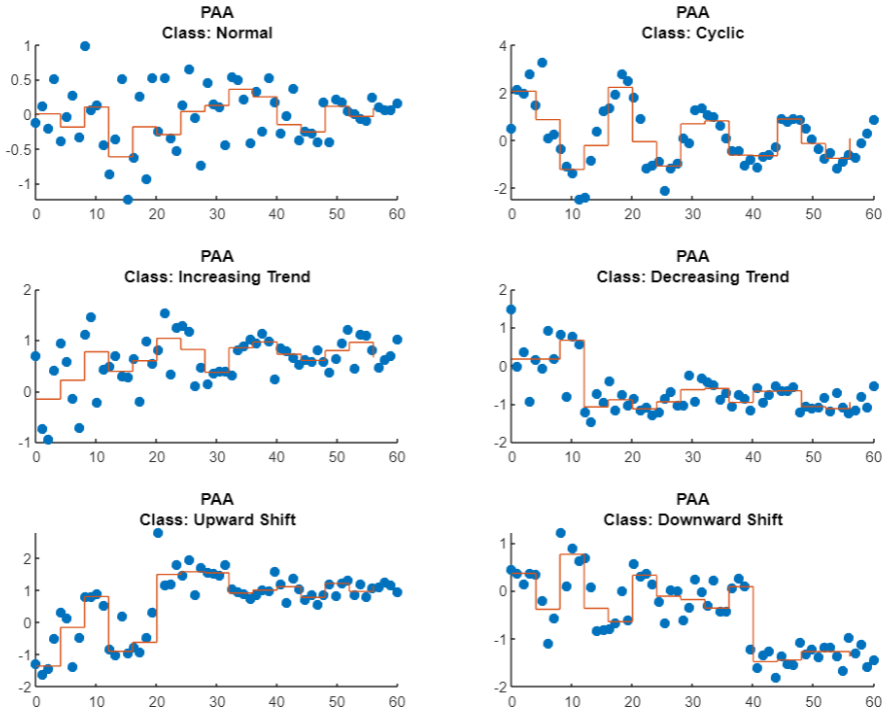
Data Confusion Matrix



It can be seen that there is not much descrepency between the Euclidean and Manhattan formulas, neither is particularly better than the other. The accuracy rating for the this data using the Euclidean distance formula is 90.1% while with the Manhattan formula it is 89.5%.

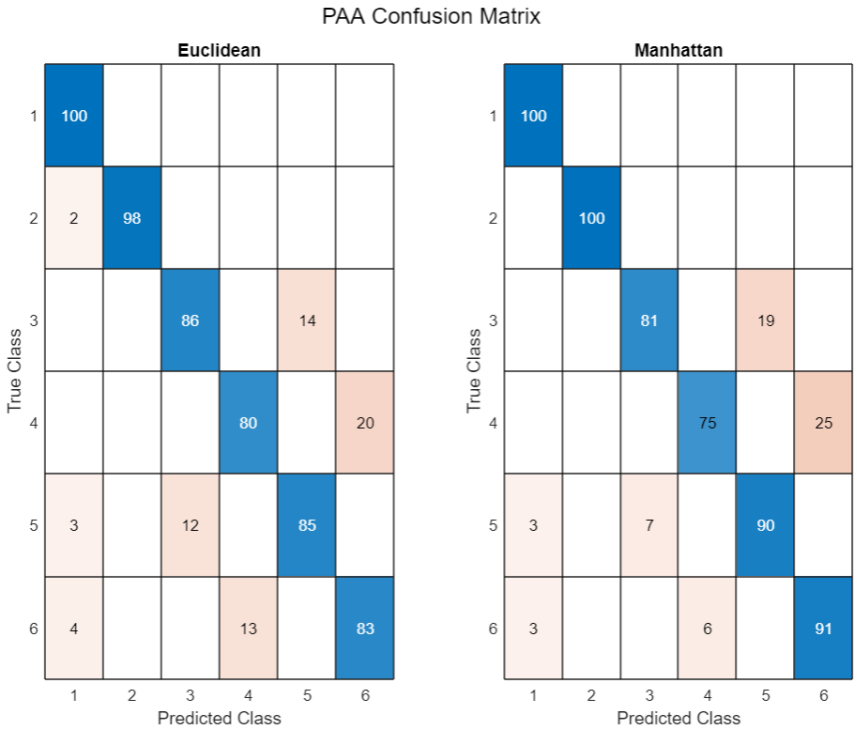
PAA

The implentaion that was used was to first determine the size of each window based on the size of the data set. The larger the window, the more accurate the generation will be, but the more processing power would be required to utilize the data. Next, the data is iterated through to determine the upper or lower bound of each window on the y-axis. The median of that sample of data is then taken to determine the output data point.
The output plot of the data set for each class can be seen below:



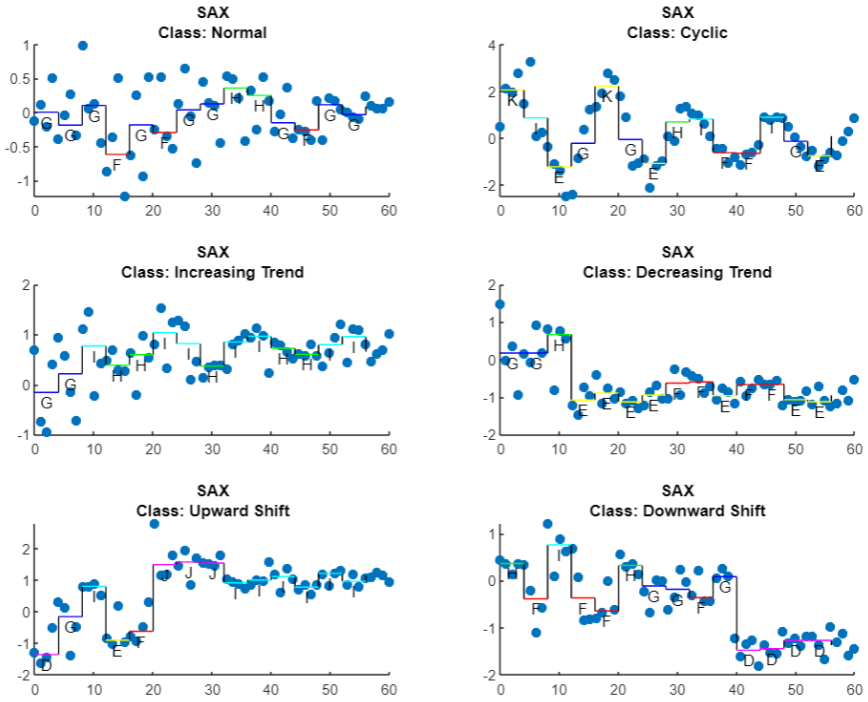
Based on the figure, it can be seen that the PAA generation tries to follow the data points as best as possible. It is impossible to have a 100% accuracy rating, but the purpose of these representaions is to get as close as possible. The accuracy rating for the original data using the Euclidean distance formula is 90.1% while the PAA genration is 88.7%. With the Manhattan formula, the raw data has a 89.5% accuracy, and with the paa generation is also 89.5%

The spread of the data can be seen in the following matrices.



SAX

The process to create the SAX was similar to implmenting the PAA, except there was an additonal element that needed to be considered, the labels. Inorder to determine the labels, the ranges of the bounds for each label needs to be defined. For this implentaion, it was chosen that for every value that could be rounded to the nearest .5 (of the normalized data) would be its own label. The labels that were picked were A-M, A starting at -3 and M ending at +3.



It can be seen in the figures above that when the horizonal lines are close enought vertically with each other they are given the same label.

Conclusion