Time Series Classification

University at Buffalo

CSE 454 - Computational Intelligence in CEN Applications

Fall 2021

Ray Chen

rchen63@buffalo.edu

Person #: 50336524

Introduction

In this project, representation and classification of time series are explored. A synthetic control dataset from the University of California Irvine is used. The dataset has 600 samples with 60 dimensions. There are 6 classes, and each class has 100 samples. Piecewise Aggregate Approximation (PAA) and Symbolic Aggregate Approximation (SAX) are the two representation techniques that are used to represent the data. K-Nearest Neighbor (KNN) is the classification technique that is used to create the classification model

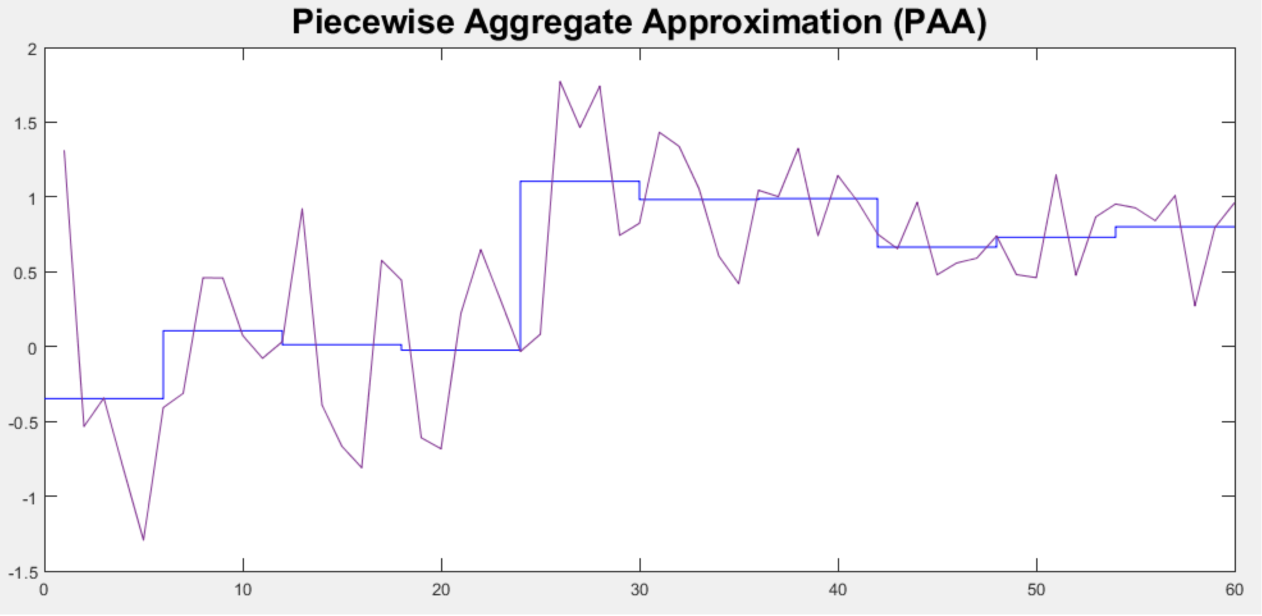
Preprocessing

The dataset standardized with a mean of 0 and standard deviation of 1.

Approach to implement PAA

1. Number of segments of PAA is selected to be 10 in this project.
2. Based on the number of segment and the dimension of the data, the length of each segment can be calculated. The length of each segment must be an integer, the length is round up if the length is not an integer.
3. The data is then separated into 10 segments. In each of the 10 segments, take the average of all the values in that segment and let the average be the value of that segment.
4. Plot PAA dataset

Plot of original time series and the PAA representation:



Approach to implement SAX

1. Use the PAA plot.
2. Create 6 alphabetic symbols for Gaussian Distribution, their ranges are:

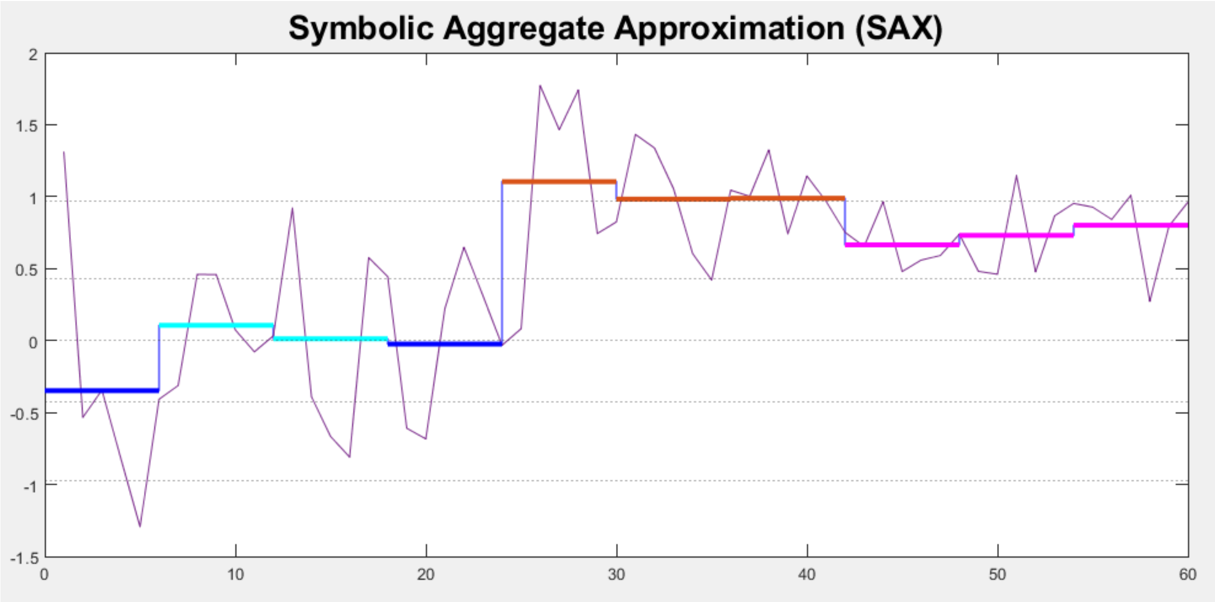
   ```
   a = [-inf, -0.97]
   b = [-0.97, -0.43]
   c = [-0.43, 0]
   d = [0, 0.43]
   e = [0.43, 0.97]
   f = [0.97, inf]
   ```

3. Label the PAA based on the range of each segment.
4. Each label has a unique color, the colors are:

   a = red
   b = green
   c = blue
   d = cyan
   e = magenta
   f = orange

5. Plot each segment of PAA with specific colors based on the label.

Plot of original time series and the SAX representation:



Establishing training and testing data generation

Training dataset and testing dataset are simply separate the origin dataset into two parts, the size ratio of training dataset to testing dataset is 7:3

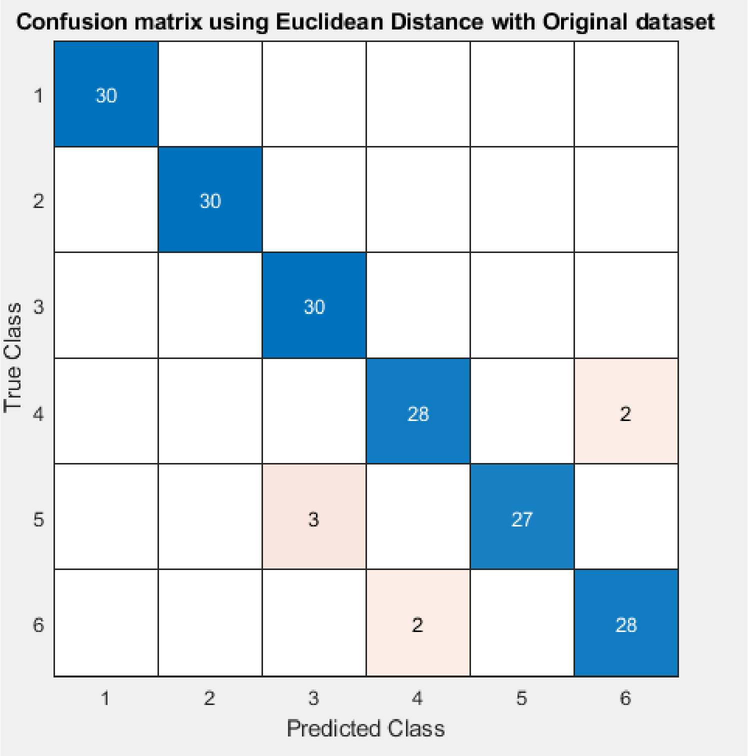
Classification process and result

fitcknn is the classifier used to train the model.
(Fit k-nearest neighbor classifier - MATLAB fitcknn (mathworks.com))

predict: Predict labels using k-nearest neighbor classification model
(https://www.mathworks.com/help/stats/classificationknn.predict.html)

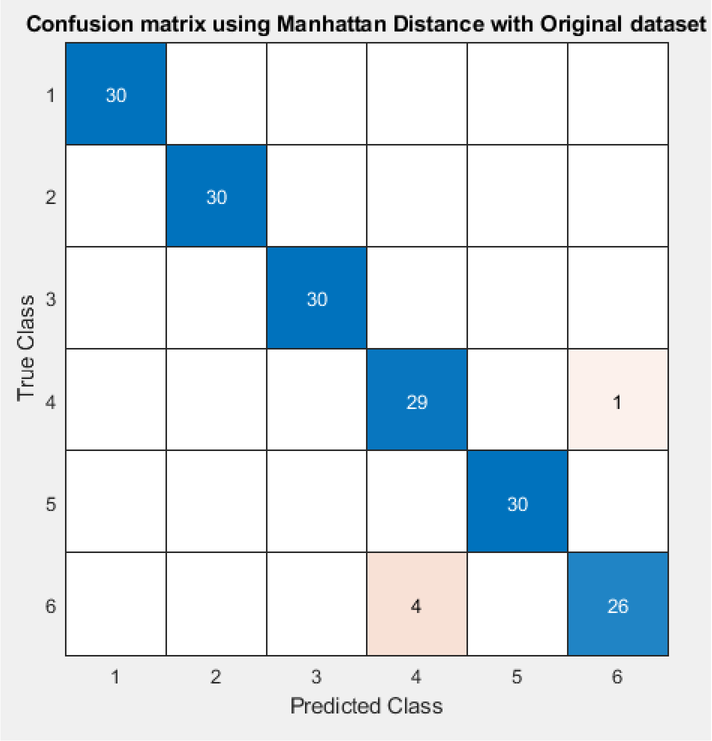
- Classification using original dataset and Euclidean distance

Accuracy = 0.9611



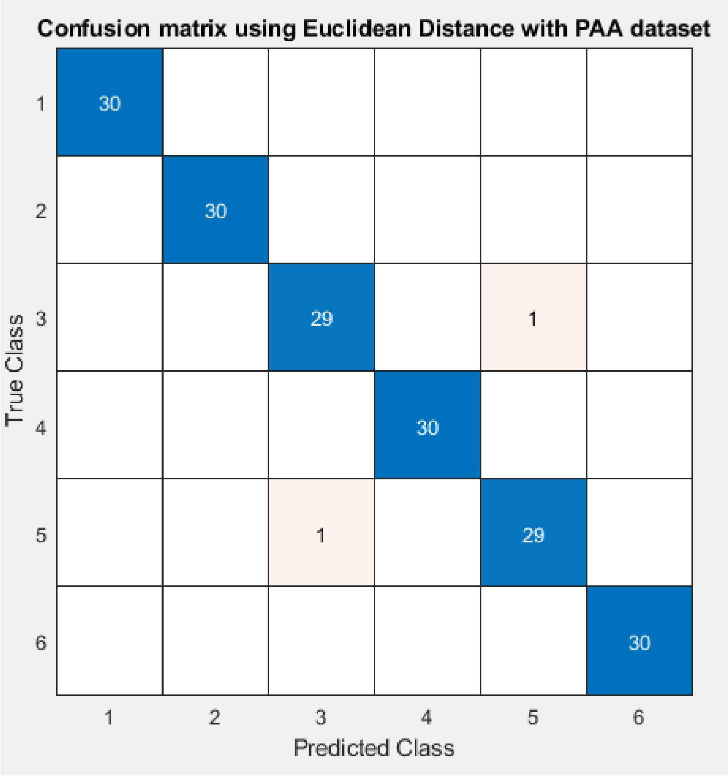
- Classification using original dataset and Manhattan distance

Accuracy = 0.9722



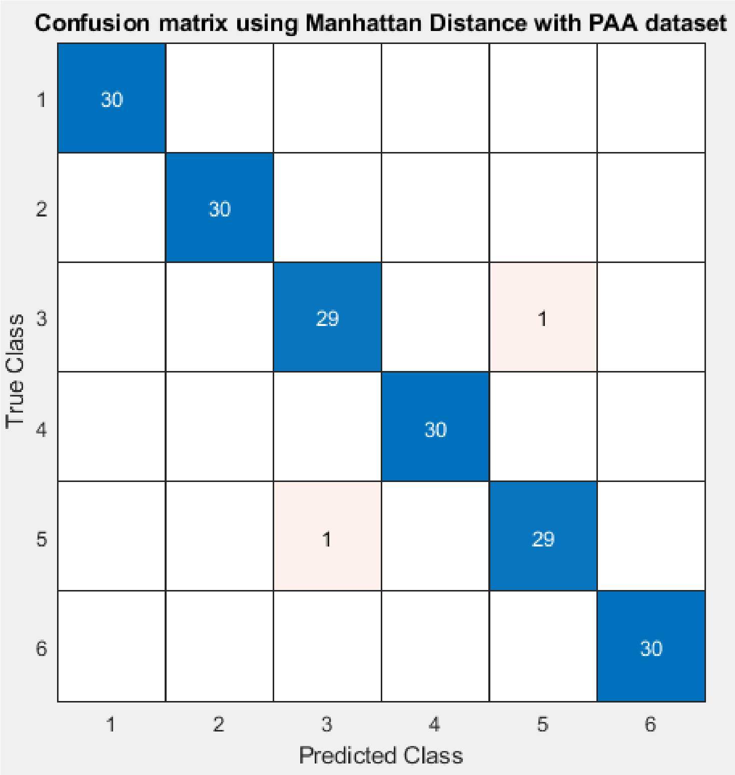
- Classification using PAA dataset and Euclidean distance

Accuracy = 0.9889



- Classification using PAA dataset and Manhattan distance

  Accuracy = 0.9889



Confusion matrix using Manhattan Distance with PAA dataset

True Class \ Predicted Class	1	2	3	4	5	6
1	30					
2		30				
3			29		1	
4				30		
5			1		29	
6						30

As we can see from the accuracies of the classification processes, classifications using PAA dataset have better accuracies than using the original dataset, furthermore, classifications using PAA dataset also use less time for training because PAA dataset has less dimensionality.

Classification using original dataset has better accuracy when using the Manhattan distance function, but in classification using PAA dataset, Euclidean distance and Manhattan Distance function output the same accuracy.

References

Lin, J., Keogh, E., Lonardi, S. & Chiu, B.
"A Symbolic Representation of Time Series, with Implications for Streaming Algorithms."
In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and
Knowledge Discovery. San Diego, CA. June 13, 2003. https://cs.gmu.edu/~jessica/sax.htm,
htps://www.cs.ucr.edu/~eamonn/SAX.htm