Data Mining – Home Assignment 2

26th Nov' 2019

Shweta Suran

• Exercise 1

To visualize the Dynamic Time Wrapping, two datasets were used; along with one dataset was made (manually). First, program load the dataset (IstanbulStockExcchange & ECG200) in CSV format or can use ManualData). Two vectors 'VectorA & VectorB' saved the two different columns of the dataset. After that program use Edit Distance to calculates the distance between points from 'VectorA & VectorB'. Then the program calculates distance among the all pair of vectors. After this program select the pairs of vectors with minimum distance for each iteration. Figure 1, 2 & 3 shows the results.

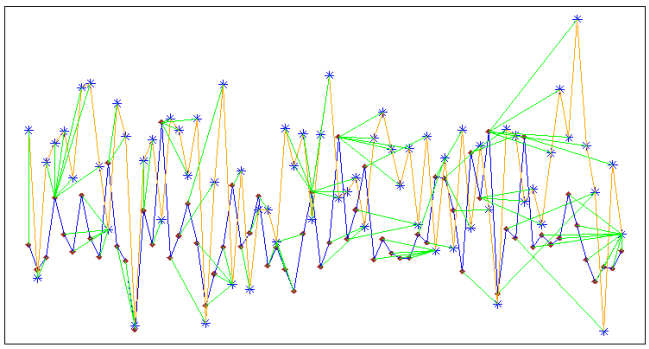

Figure 1:DTW-Istanbul Stock Exchange

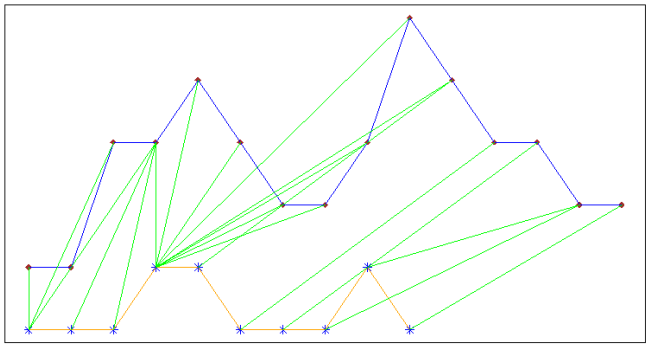

Figure 2: DTW-Manual Data

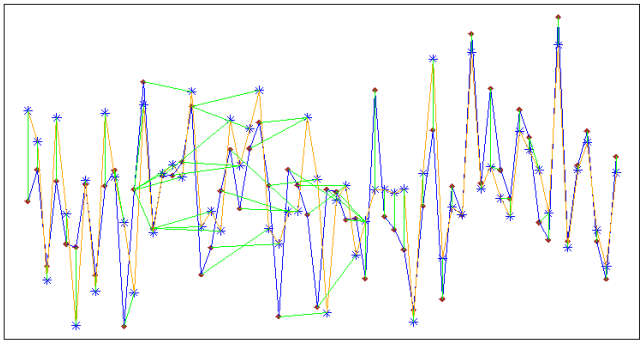

Figure 3: DTW-ECG200

• Exercise 2

Program generates 'VectorX & VectorY' of length 400 by using 'rnorm' fucntion. Then the program calculated the K-NN (using Euclidean distance function); k=3. In next step, program calculates the LRD (Local Reachability Density) and then calculates the LOF (Local Outlier Factor). The Figure 4 shows the original data and Figure 5 shows LOF.

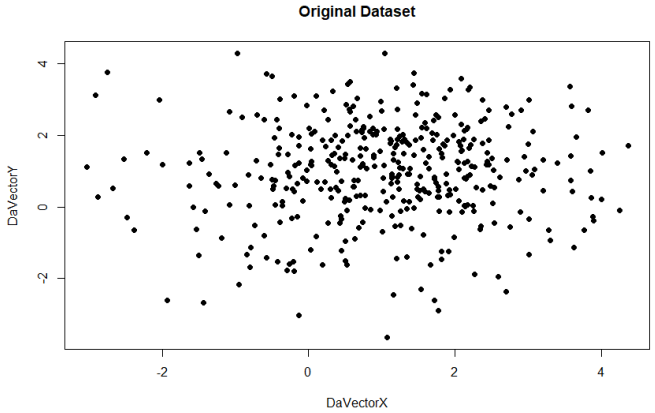

Figure 4:Original Data

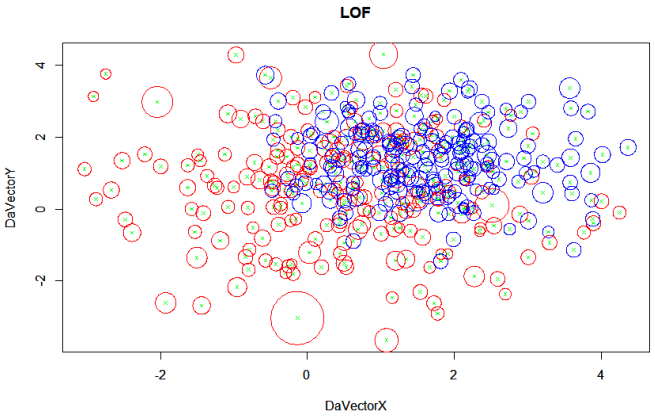

Figure 5: Local Outlier Factor (LOF)

• Exercise 3

Program generates the data stream of 750000 data points. Then, 500 points (initial) store into reservoir in 'PVectorX & PVectorY'. After this, program go across the other (remaining) data points in such a way that the probability of each point computed as 500/point (stream indx). Figure 6 shown the Real Reservoir, Figure 7, 8 & 9 show data streaming at different indexes (data points at 7500, 175000 & 750000 iterations) and finally Figure 10 shows the number of changes in the reservoir.

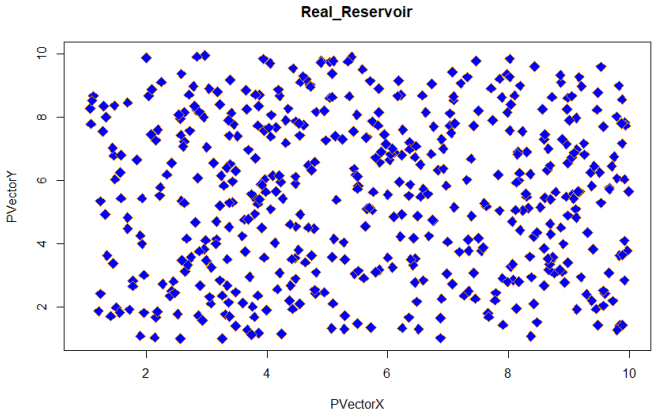

Figure 6: First Reservoir

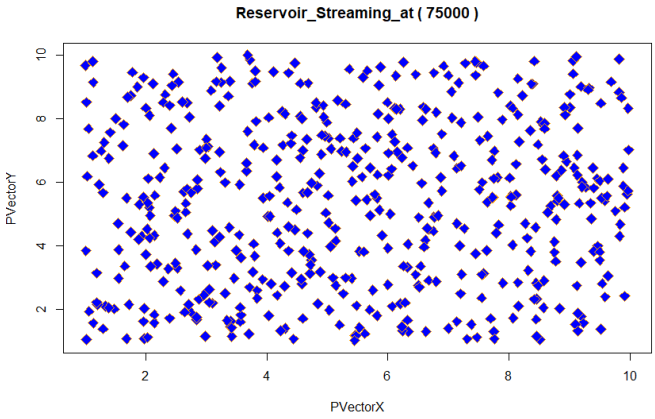

Figure 7: Reservoir Sampling at Index 75000

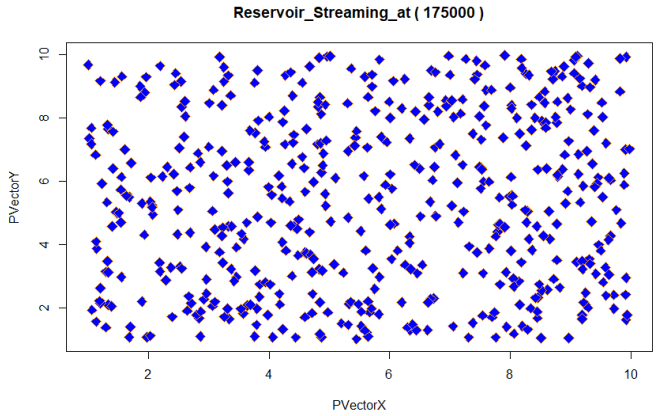

Figure 8: Reservoir Sampling at Index 175000

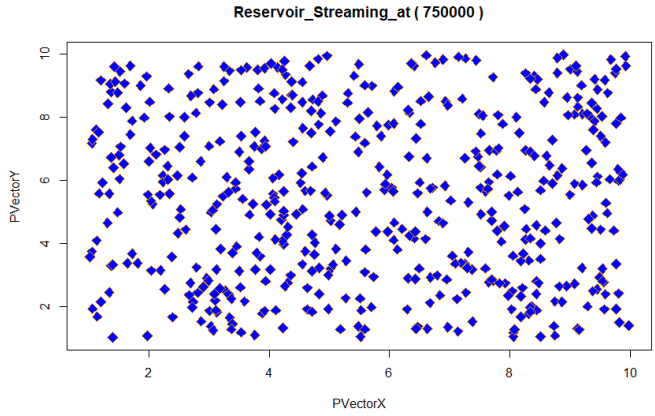

Figure 9: Reservoir Sampling at Index 750000

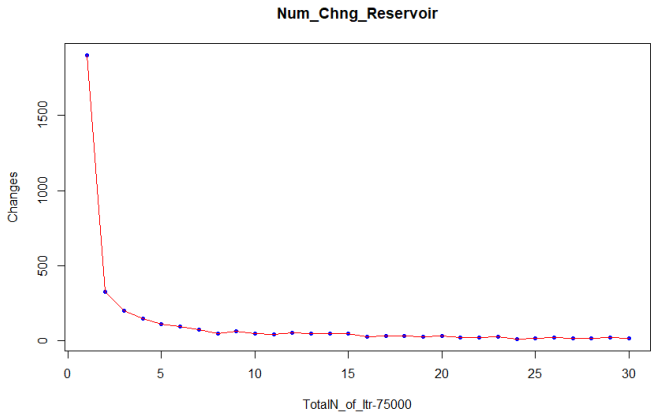

Figure 10: Number of Changes in the Reservoir

Exercise 4

First program imports the data and then computes the moving average by calling the function 'MovAverage1' and one more time computes the moving average from the output by calling the function 'MovAverage2'. After this, YmAv (results) are divided by mAv and then average for each quarter is computed. Output values are used to obtain Deseasonalized data and then it passed through the filter. Then coefficients are calculated which further used to compute the forecasting. The Figure 11 shows the original, trend and forecasting data.

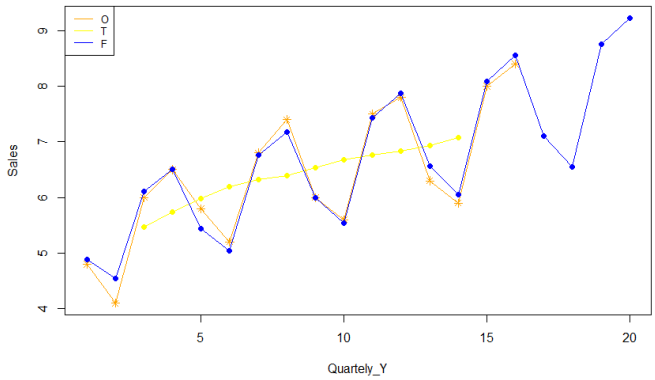

Figure 11: Forecasting (Original: Orange, Trend: Yellow, Blue: Forecast)

• References

- LOF Detection- https://www.datasciencecentral.com/profiles/blogs/anamoly-outlier-detection-using-local-outlier-factors
- How DTW (Dynamic Time Warping) algorithm works- https://www.youtube.com/watch?v=_K1OsqCicBY
- Tutorials Point-https://www.tutorialspoint.com/
- Euclidean Distance- https://stackoverflow.com/questions/45780199/function-to-calculate-euclidean-distance-in-r
- Tutorial Gateway-https://www.tutorialgateway.org
- Reservoir Sampling- https://www.geeksforgeeks.org/reservoir-sampling/
- Time Series Forecasting- https://www.youtube.com/watch?v=gHdYEZA50KE https://www.youtube.com/watch?v=mC1ARrtkObc