Master Thesis

Unsupervised Detection of Patterns from Time Series Data

Rakesh Lagare

Media Informatics

Advisors : Shreekantha Devasya, Alexander Graß

Supervisors : Prof. Dr. Jarke, Prof. Dr. Beecks

Fraunhofer FIT

Overview

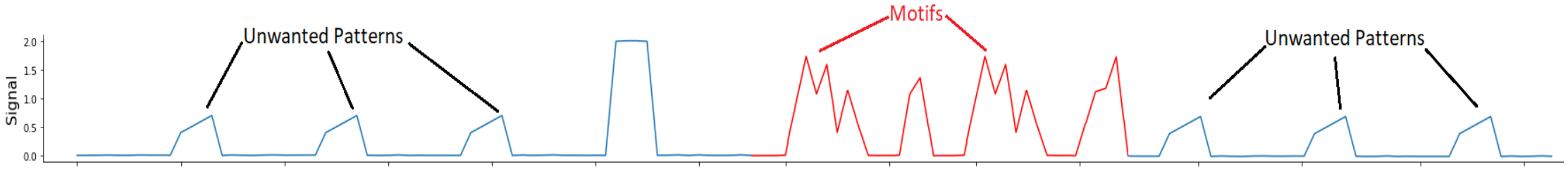
- Terminologies
- Motivation
- Goal
- Existing Methods
- Implementation
- Evaluation
- Conclusion

Terminologies

- A **time series** is a sequence of observations measured at certain time intervals.
- **Motifs and Discords** are one of the influential representative patterns in time series.
- **Motifs** are frequently occurring repeated patterns of a time series.
- **Discords** are outliers of a time series.
- Motif discovery has been used in different aspects, such as, telecommunications, medicine, web and sensor networks.
- In the last couple of decades lots of approaches have been proposed to address similarity search issue in time series data.

Motivation/Idea

- Huge Data was collected by MONSOON EU project , which was shared by sectors of aluminium and plastic.
- Since the data was huge and multi-dimensional , it was difficult to analyse the complete data.
- In that scenario Automatic discovery of Motifs and Discords would have been helpful.
- **Hence a more general and semi-automatic approach is proposed**.

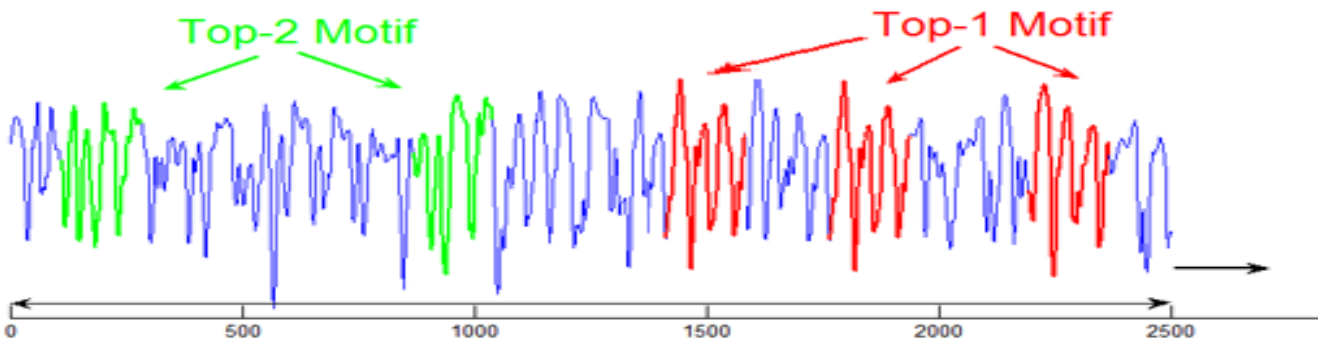


Fraunhofer FIT

Goal

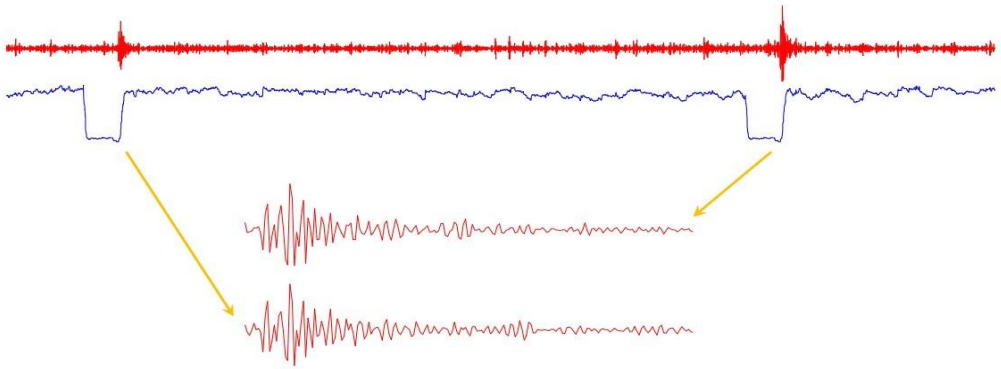
- An efficient way of performing semi automatic similarity search that identifies **motifs** as well as **discords** and rank them based on frequency of occurrences.

- Proposed approach is compared against the existing state of the art approach which demonstrate its quality improvement, provide fast and precise similarity search.



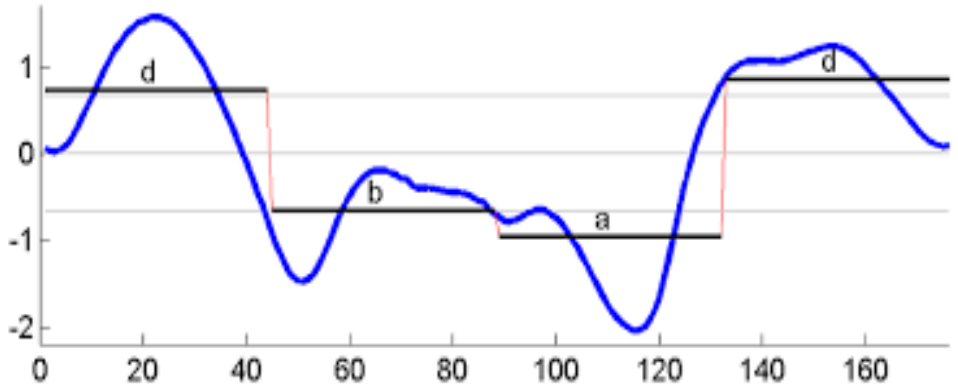
Existing Methods : Matrix Profile

- Takes an arbitrarily long, unlabelled time series and produces intuitive and most matched patterns.
- It generally does not attempt to explain all the data in the time series, rather only considers salient sub-sequences.
- Minimum Description Length (MDL) is used to extract such kind of sub-sequences.
- **Drawback:** Since it uses Nearest Neighbours method, chances of missing all other important patterns are very high.



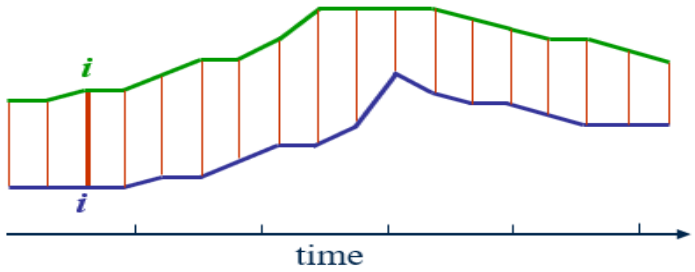
Existing Methods : Symbolic Aggregate Approximation (SAX)

- In SAX, time series data transformed into equal sized segments.
- Each segment is symbolized into a sequence of strings.
- It minimizes dimensionality by the mean values of equal sized frames.
- **Advantage:** Provides low computational complexity.
- **Drawback:** This mean value based representation causes a high possibility to miss important patterns.

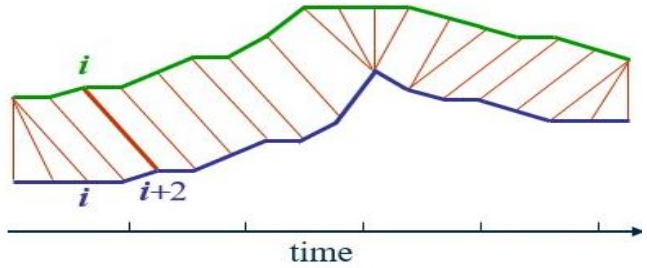


Existing Methods : Exhaustive Search using Dynamic Time Warping (DTW)

- Algorithm to measure similarity between two time series which may vary in time or speed.
- Sequence of techniques are used to fasten up similarity search to find out motifs in a time series.



Euclidean distance aligns the i-th point on one time series with the i-th point on the other.



A non-linear(DTW) alignment produces a more intuitive similarity measure.

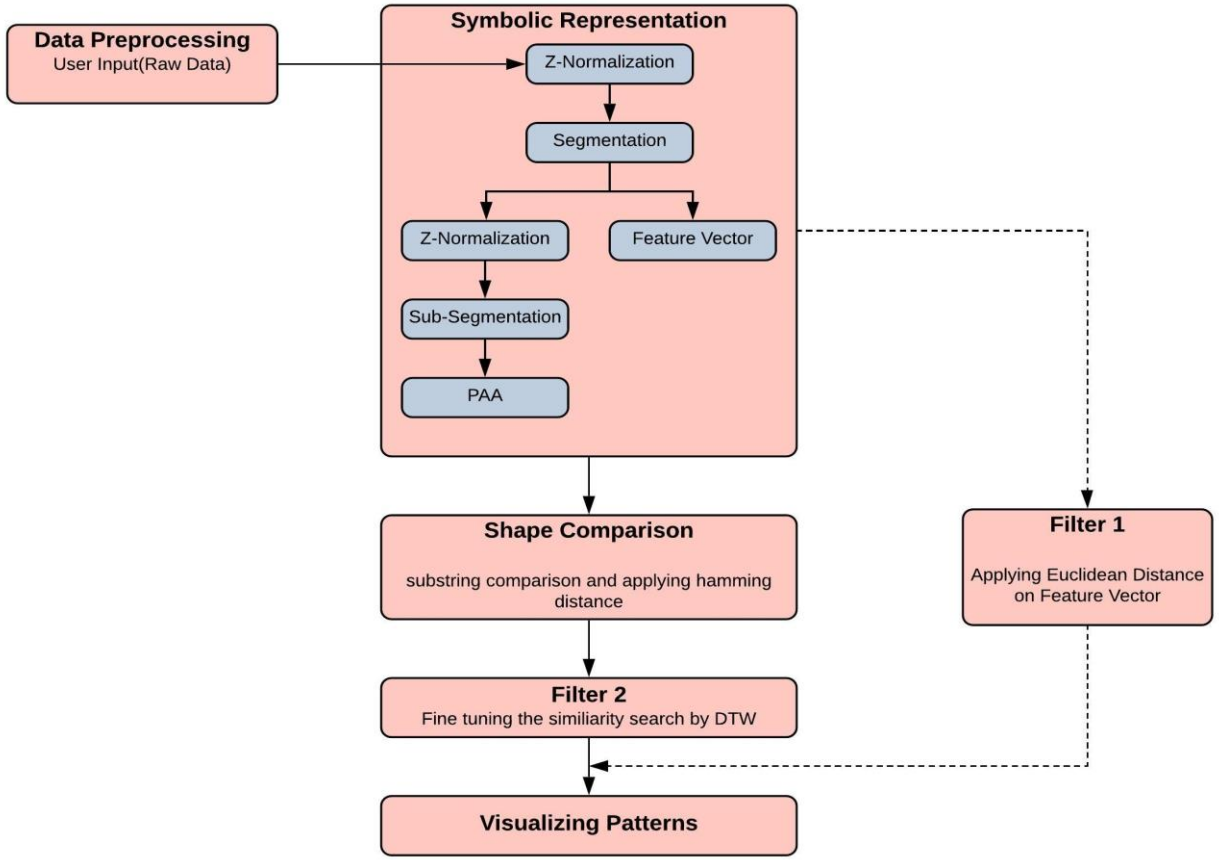
Drawback:

- Despite the effectiveness of the DTW, it has an **O(M.N^2)** time and space complexity .

Source https://www.psb.ugent.be/cbd/papers/gentxwarper/DTWAlgorithm

Implementation

Proposed multi-step approach is achieved by 6 sub steps.

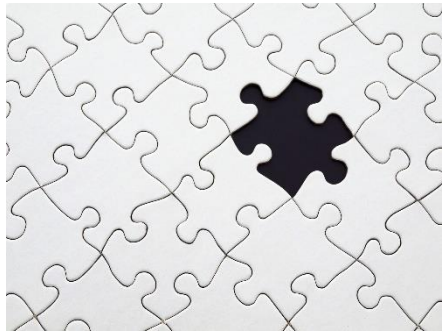


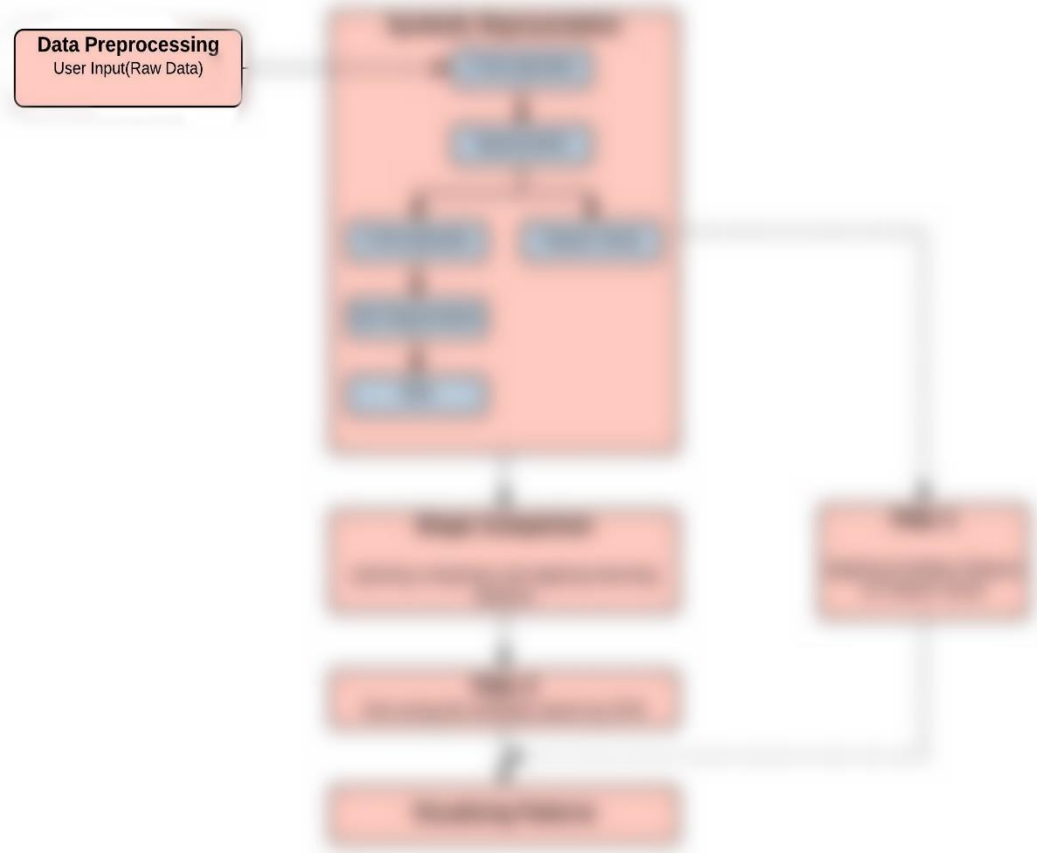
Fraunhofer FIT

Implementation

Data Pre-processing

- Time series dataset is provided as a input file.
- Data is pre-processed to remove unwanted values and impute the missing values.



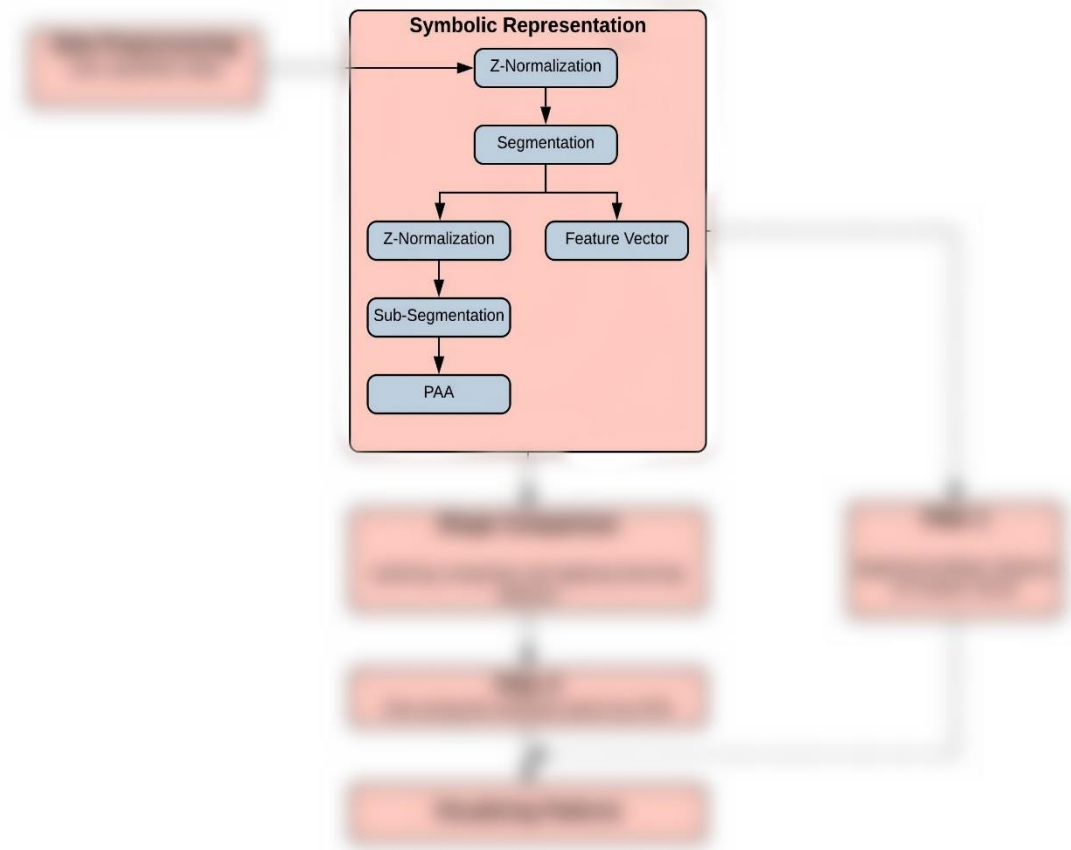


Fraunhofer FIT

Implementation

Z-Normalization

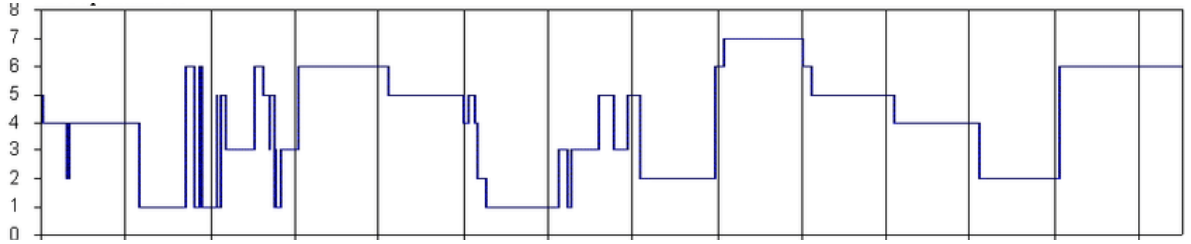
- Now Pre-processed data is z-normalized.
- Process of normalizing the data to the mean value is 0 and the standard deviation is approximately 1.
- The reason behind this Z-normalizing is to scale down all the patterns to the same level.

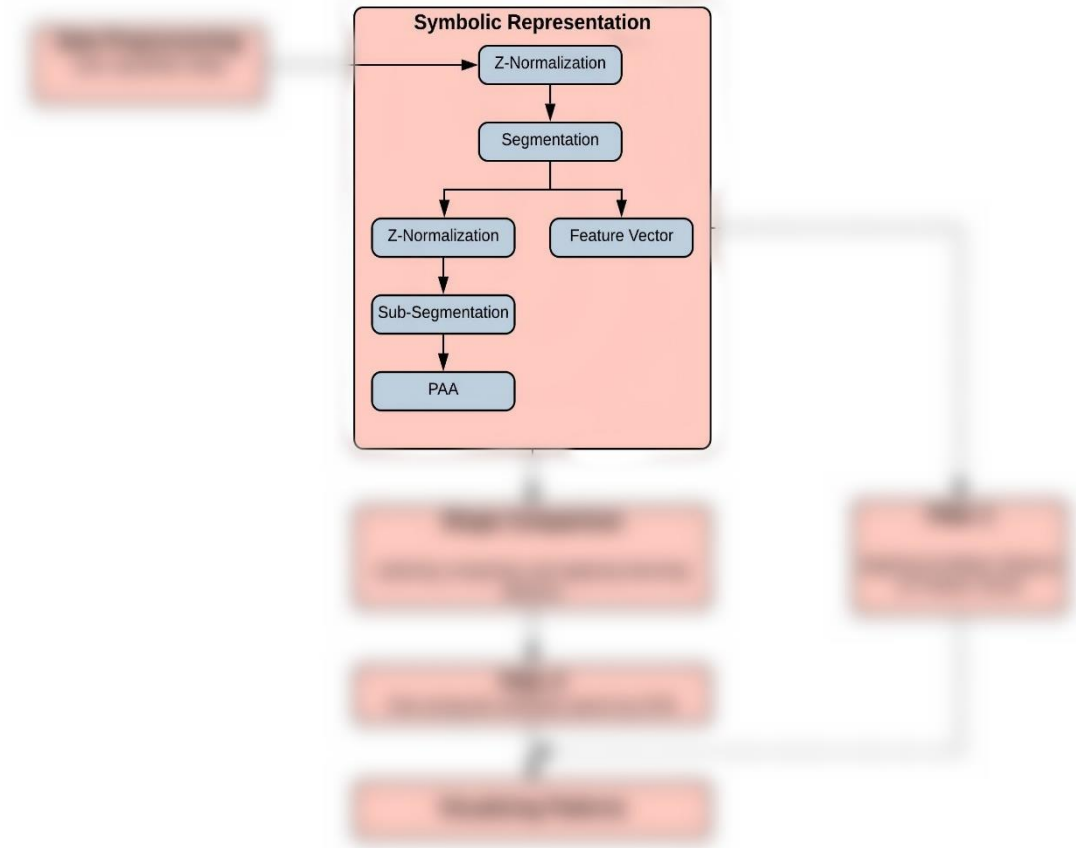


Implementation

Segmentation

- Normalized data is segmented into smaller sequences each of length ***n*** using fixed window.
- Sliding window can be used optionally.
- All extracted segments data is stored for further

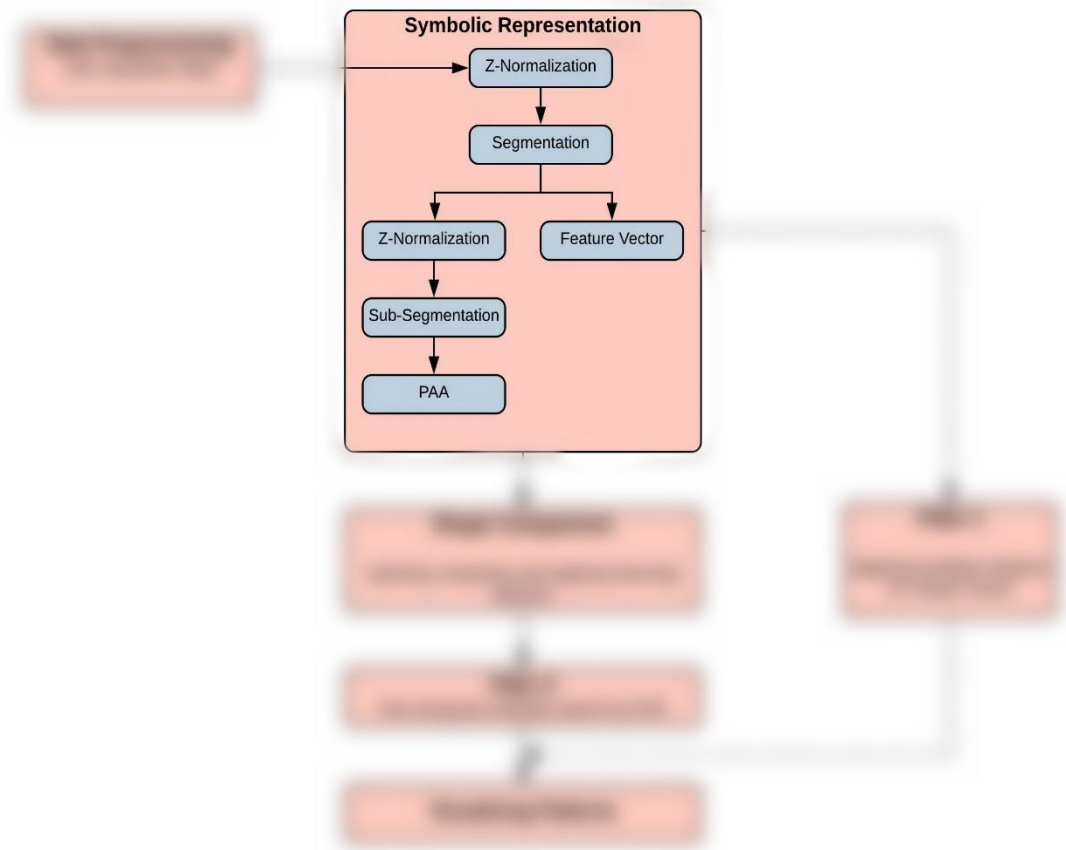




Implementation

Feature Vector

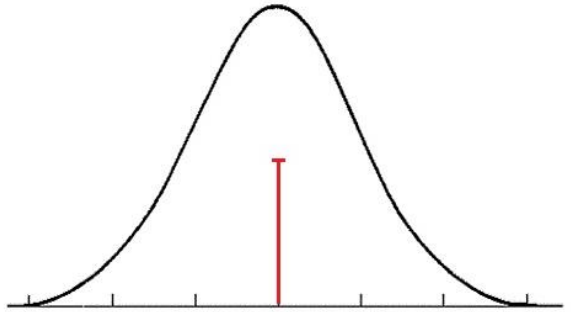
- Segments are again normalized hence we lose some important features of these sequences.
- To store features of these sub-sequences in the feature Vector.
- Features : Scale , Offset/Position



Implementation

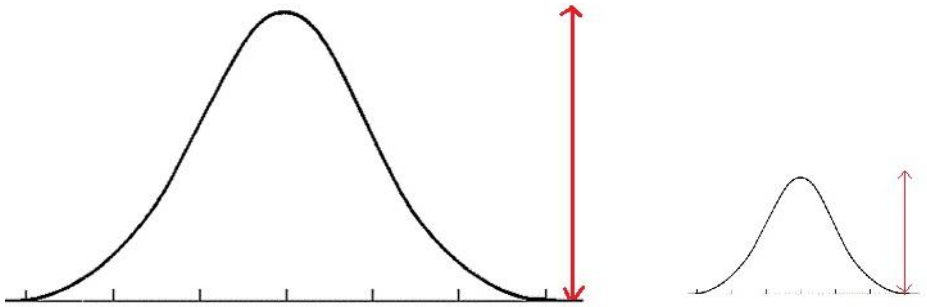
Offset

For each segment, mean value of data points is calculated.



Scale

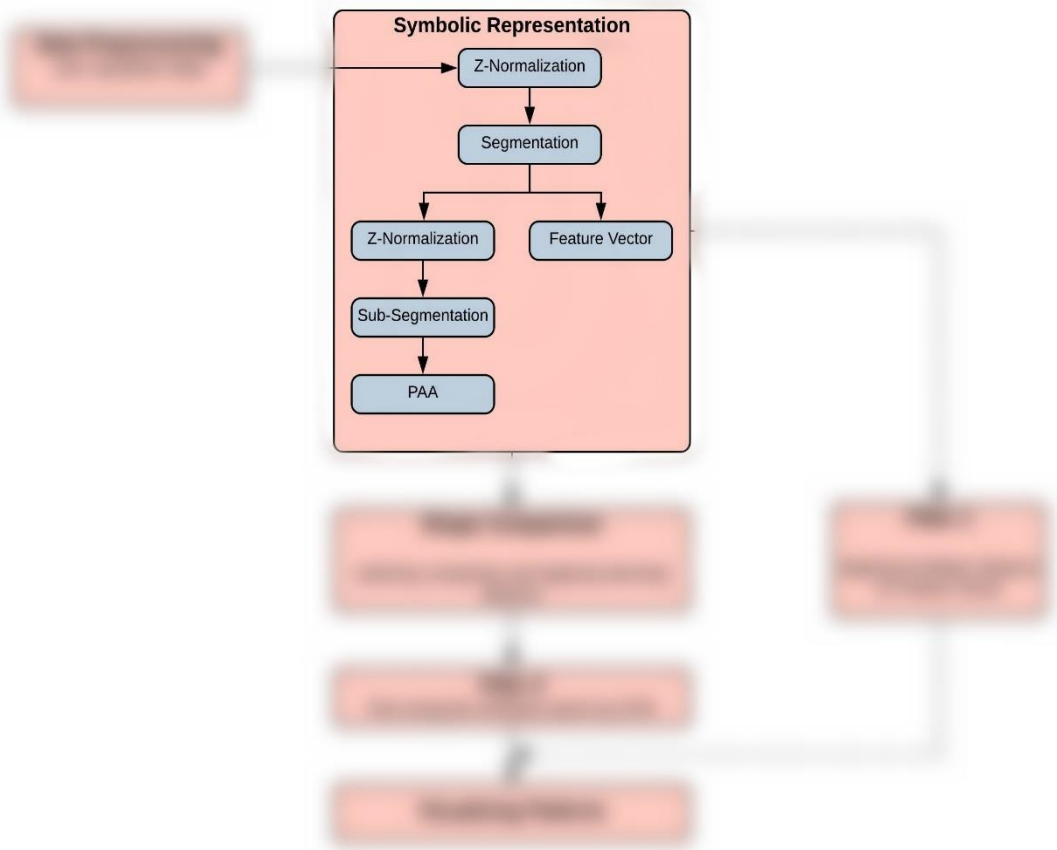
For each segment ,subtracted value of highest and lowest data points.



Implementation

Segment Z-Normalization

- Now segments are again Z-normalized.
- To pick an equivalent region under the Gaussian curve by utilizing look up table for cutting y -co-ordinates.



Sub-Segmentation

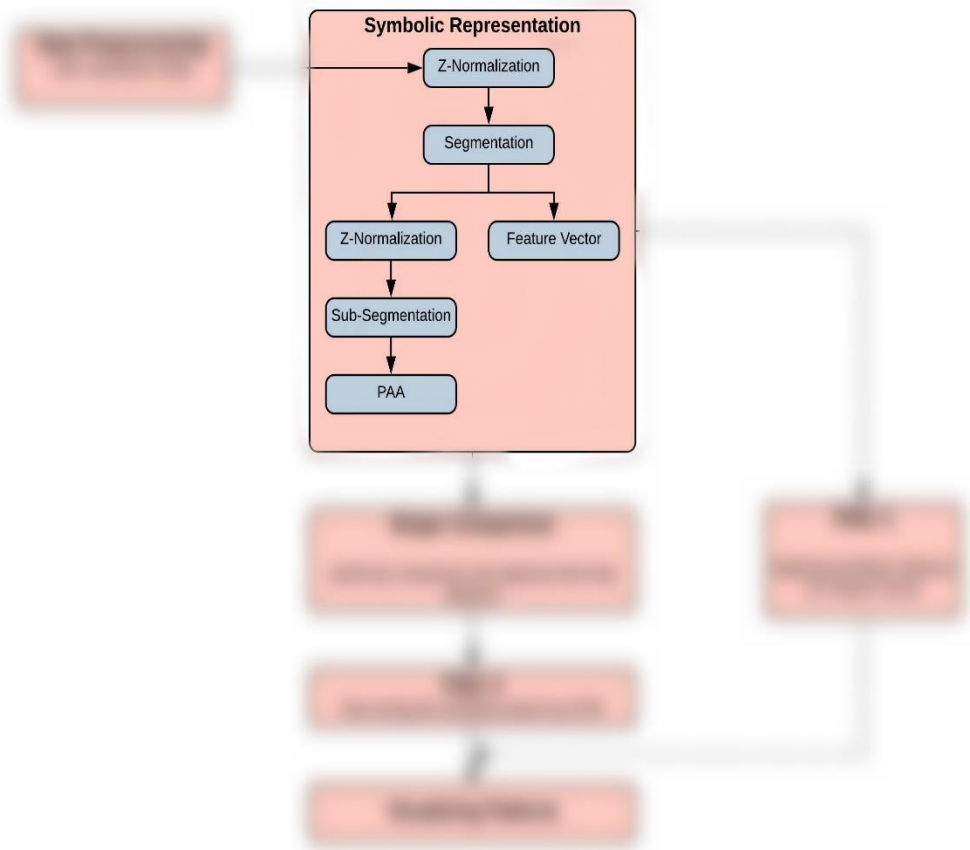
- To avoid missing out any patterns, segmented data again segmented into smaller sub segments of x co-ordinate word length "s".

Implementation

Piecewise Aggregate Approximation (PAA)

- Used for dimensionality reduction.
- Area under the Gaussian curve of sub-segment are **break points**.
- Mean value of each Sub-segment data is called **PAA coefficients**.
- PAA coefficients are changed into letters by using a look-up table.

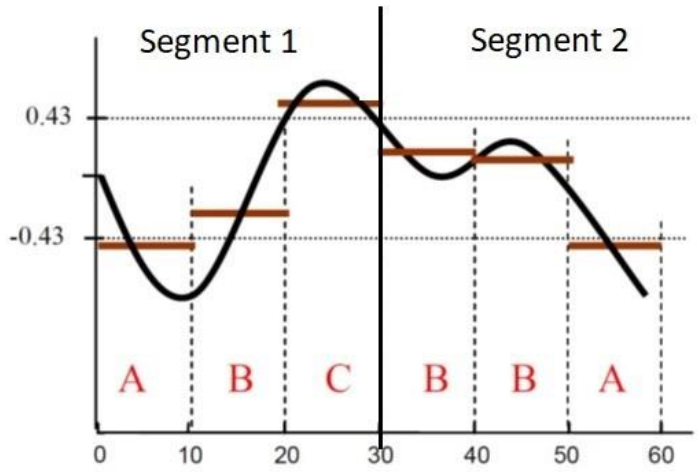
	3	4	5	Letter
β1	-0.43	-0.67	-0.84	A
β2	0.43	0	-0.25	B
β3		0.67	0.25	C
β4			0.84	D



Source: A symbolic representation of time series, with implications for streaming algorithms

Implementation

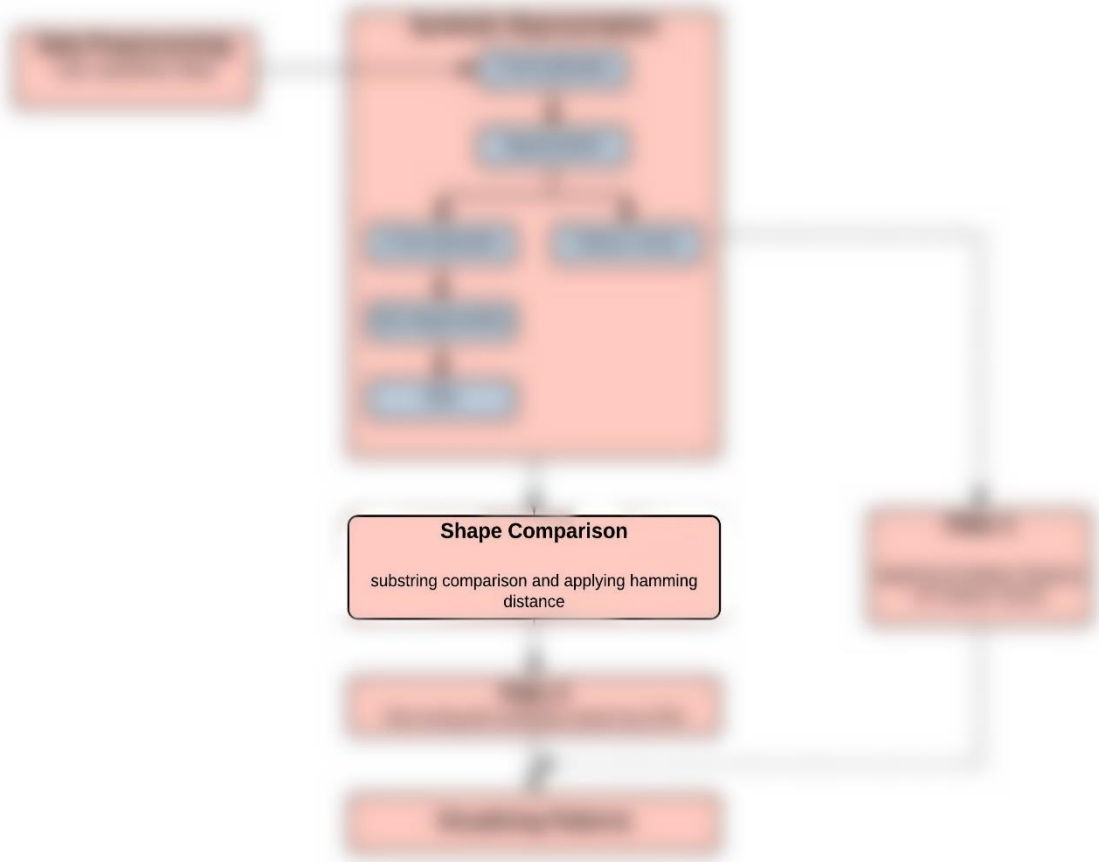
- All PAA coefficients below the smallest break point are mapped to the symbol "A“
- All coefficients greater than or equal to the smallest break point and less than the second smallest break point are mapped to the symbol "B" , so on.
- Example :
  - length of the time series (n) = 60,
  - no. of segments (k) = 2 (0-30, 30-60)
  - word size (a) = 3, so sub-segments =3
  - segment 1 mapped to ABC , segment 2 mapped to BBA.



Implementation

Shape Comparison Algorithm

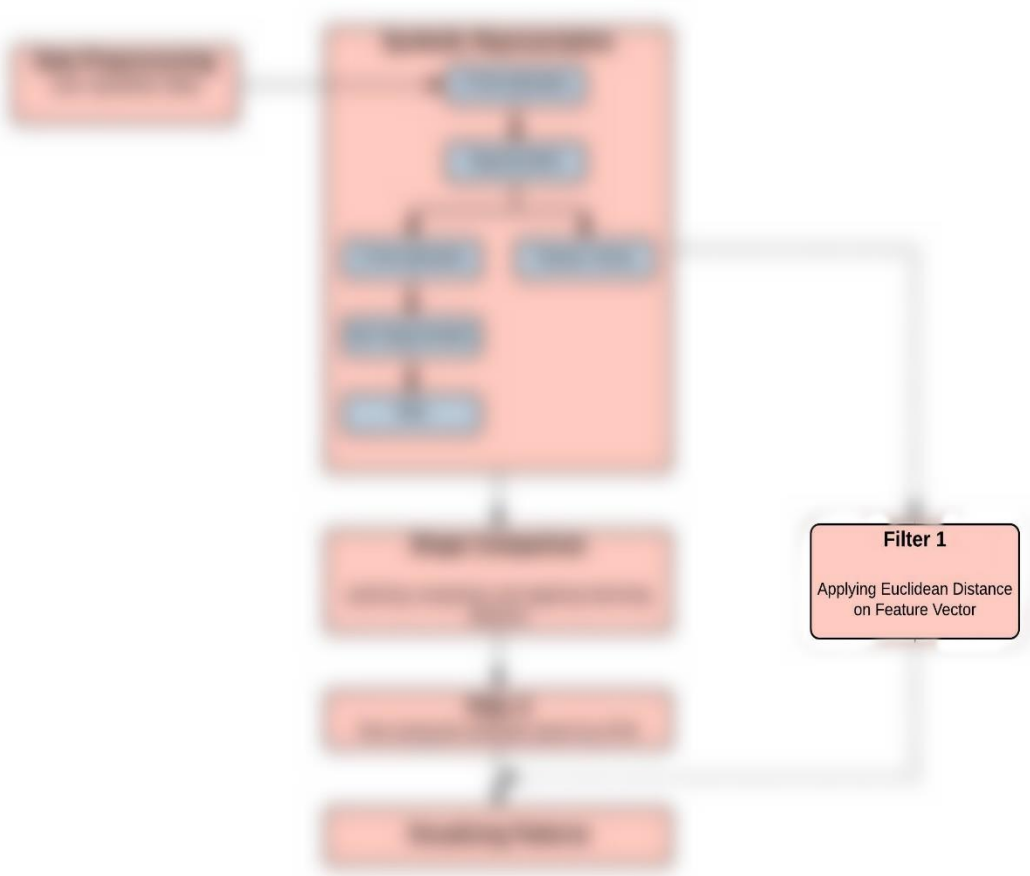
- Each string is compared against other remaining strings .
- Matched strings are grouped and stored in matrix along with their indices.
- Theoretically, all the members in each group will be having the same shape.
- Optionally, **Hamming Distance** is used to avoid the loss of true positive patterns.



Implementation

Filter 1: Feature comparison by Euclidean Distance

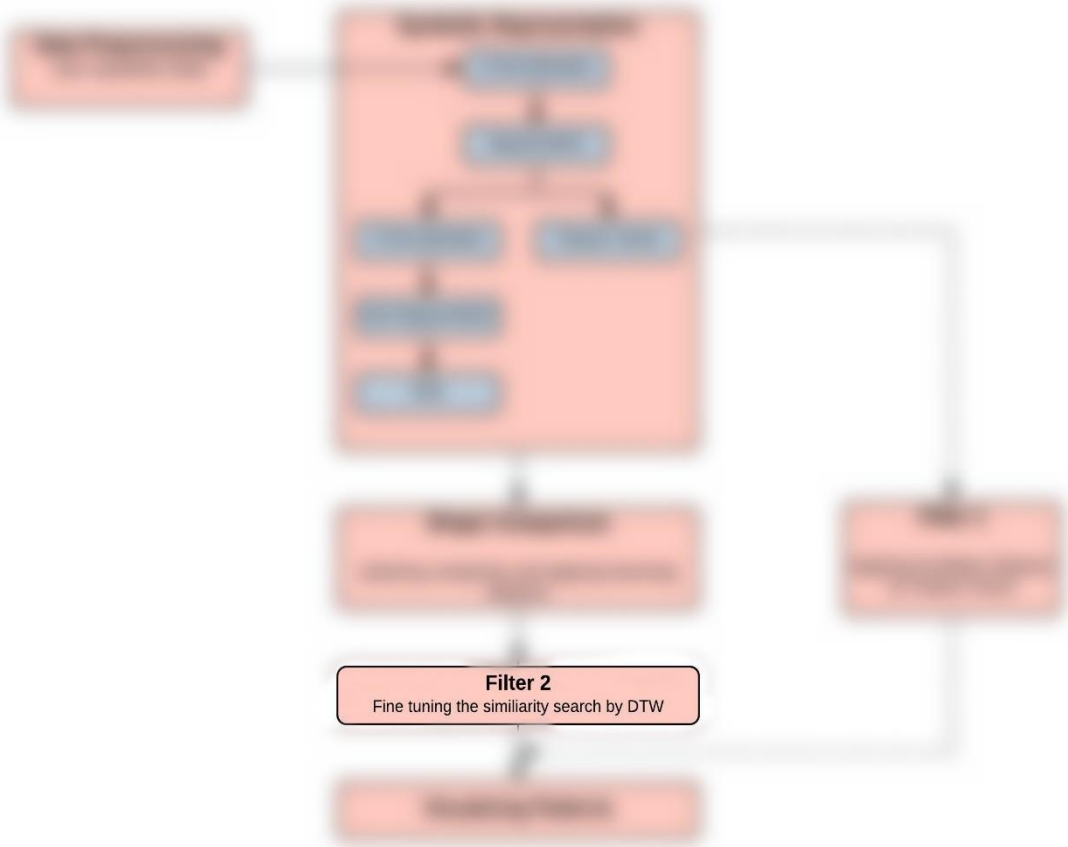
- Optionally using **scale** or **offset** from feature vector to provide more accurate results such as sub class categorisation.
- Feature vector is proposed to achieve both accuracy and efficiency.
- First locates the patterns along the transformed data based on the strings.
- Later it is used to filter out the patterns based on features.



Implementation

Filter 2: Fine tuning the similarity search using Dynamic Time Warping

- Applying DTW for results to further narrow down and provide precise patterns.
- Rank table is prepared according to DTW value and the frequency of occurrences.
- All the patterns obtained from the previous step are visualized with respect to the rank table.



Evaluation

Accuracy:

- Synthetic data - **Normalized and Non-Normalized** datasets are prepared.
- Evaluator knows all the patterns and has the better understanding of the patterns.
- Using brute force algorithm, Sub-sequence are captured.
- Class and Sub-class categorization are used as metrics.

Performance: Execution time is evaluated.

No	Dataset	Length	Description
1	UCR	15-6000	UCR Time Series Data
2	ECG	120	Electrocardiogram data
3	Synthetic Data	1000 - 100,000	Randomly generated synthetic data

Evaluation

Synthetic Datasets:

Normalized Dataset :

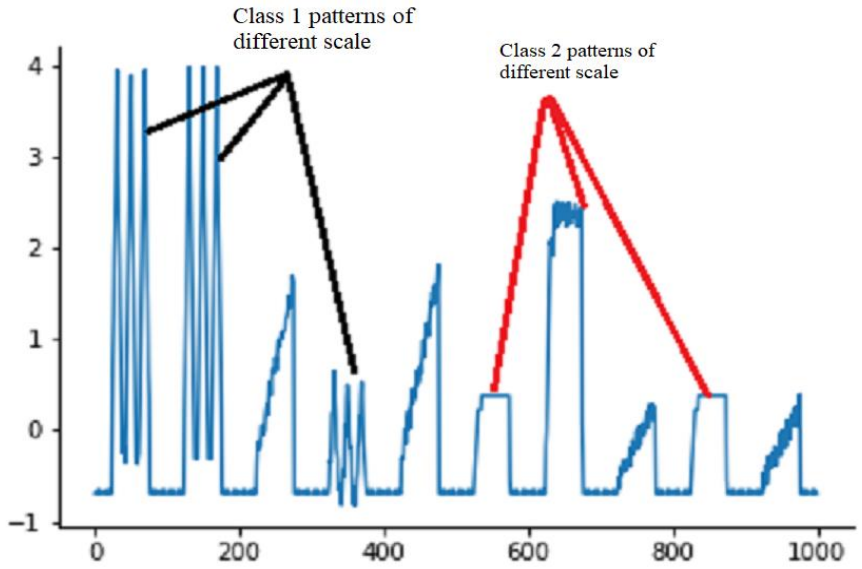
- Includes only Noise but no Scale and Offset.
- Depending on the random number generated , pattern data is generated.
- If generated random noise is 1 , noise is added to the above data.
- For class category evaluation.

Non-Normalized Dataset:

- Includes Noise, Scale and Offset.
- Depending on the random number generated , pattern data is generated.
- Depending on random numbers noise, Scale or offset added.
- For Sub-class category evaluation.

Evaluation:

- The complete dataset of size 'x' is generated, each pattern and their respective class is stored.



Evaluation: Accuracy Evaluation

Similarity Search on Synthetic Data:

- Find top-k patterns from the large set of patterns using $^{n}C_{2}$
  - n - total number of patterns/segments belong to particular class.
  - C - the class for which patterns belong.
- Similarity distance is calculated and ordered them in ascending.

Evaluation: Accuracy Evaluation

For Exhaustive Pattern Search Using DTW:

- Top 'k' sub-sequences with lowest DTW values are compared against the synthetic data table.
- Matched pairs are true positives, else false positives.
- Remaining sub-sequences outside the top-k are unwanted comparisons.

For Proposed approach :

- Each sub-sequences classified to their classes or sub-classes.
- Sub-sequences compared against synthetic data pair, matched segments are true positives ,else false positives.
- Sub-sequences outside top-k proposed approach matrix are unwanted comparisons.

Evaluation: Accuracy Evaluation

- Evaluated on 6 synthetic datasets of different size .
- True positives (pattern classification) for proposed approach is same as for EPS-DTW.
- But it has high time complexity.

Dataset Size	Top K Value	True Positives DTW	True Positives Proposed
1,000	18	18	18
2,000	43	43	43
5,000	301	299	299
10,000	1226	1,217	**1,218**
25,000	7,773	7,757	**7,758**
100,000	62,693	62,663	**62,669**

Evaluation: Accuracy Evaluation

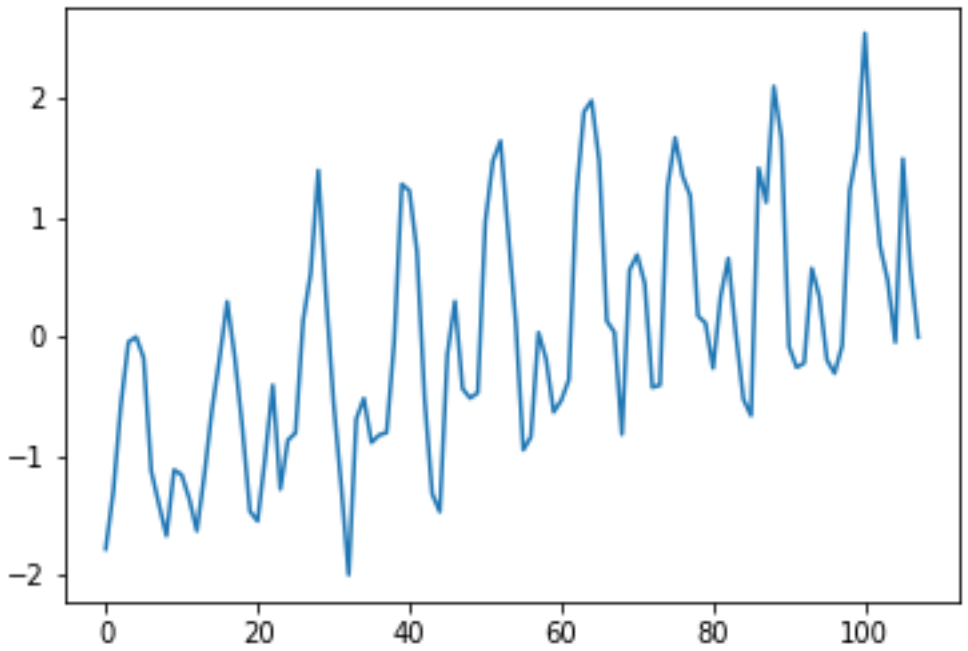
- False positives are significantly lower for proposed approach.
- Also has the lower unwanted patterns as compared to EPS-DTW.
- EPS-DTW produces more patterns than proposed approach which may not be useful or overlapping patterns.
- Results in high time complexity

Dataset Size	False Positives DTW	False Positives Proposed
1,000	0	0
2,000	0	0
5,000	6	**2**
10,000	19	**8**
25,000	61	**15**
100,000	150	**46**

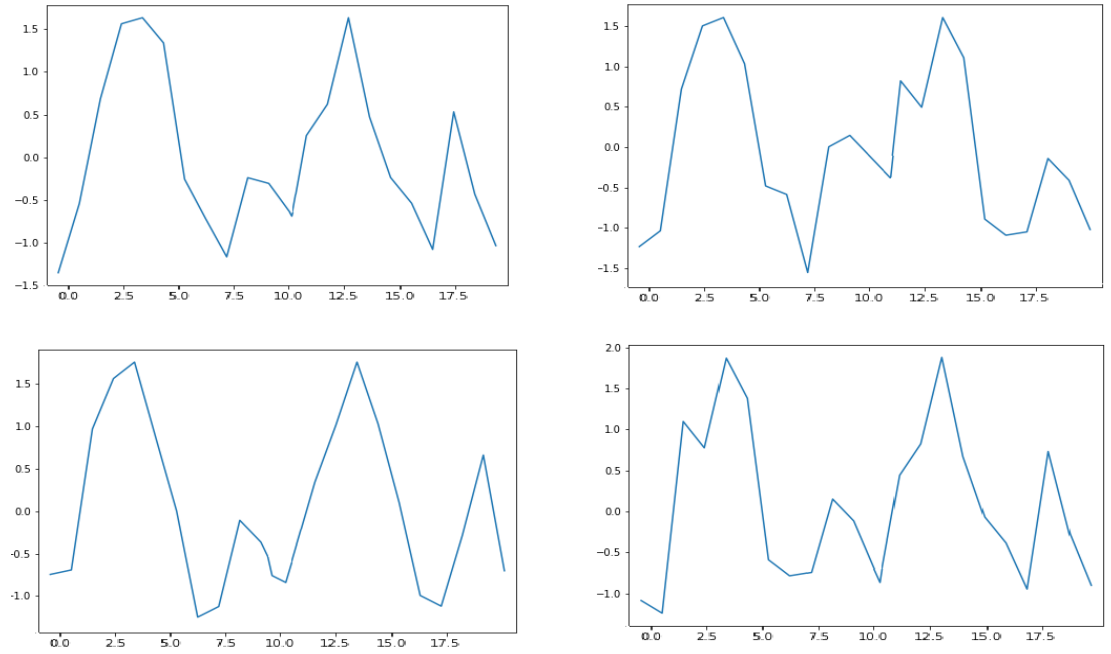
Evaluation: Accuracy Evaluation

ECG Dataset:

Original Plot:



Classified Plot: CAB



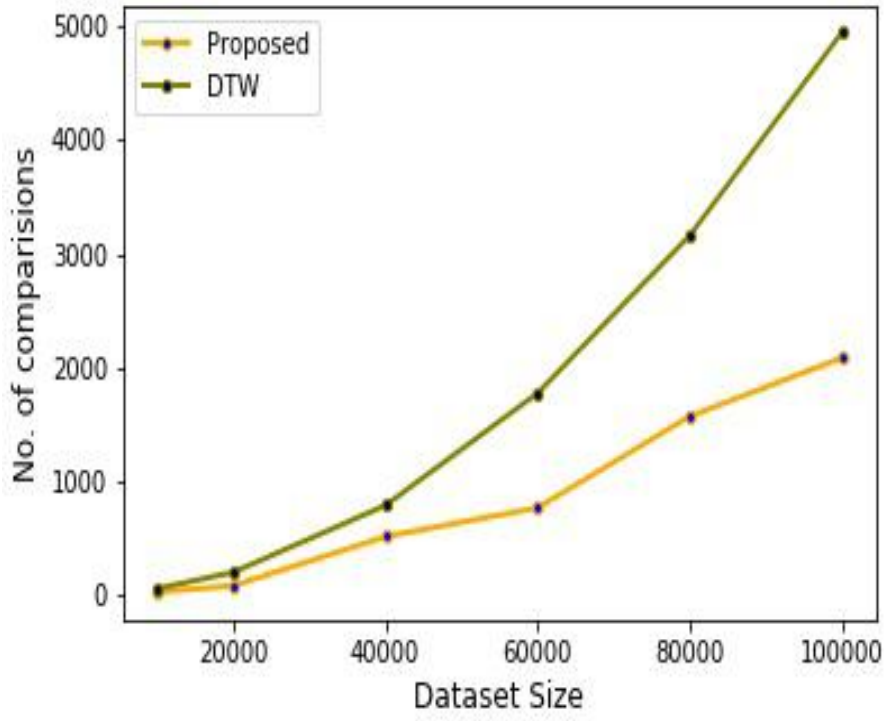
Fraunhofer FIT

Evaluation: Performance Evaluation

- Carried out on a 64 bit machine with Windows OS,
- 2.20GHz core i7 processor and an 8 GB RAM system.
- segment size (w) is kept same.

Comparisons :

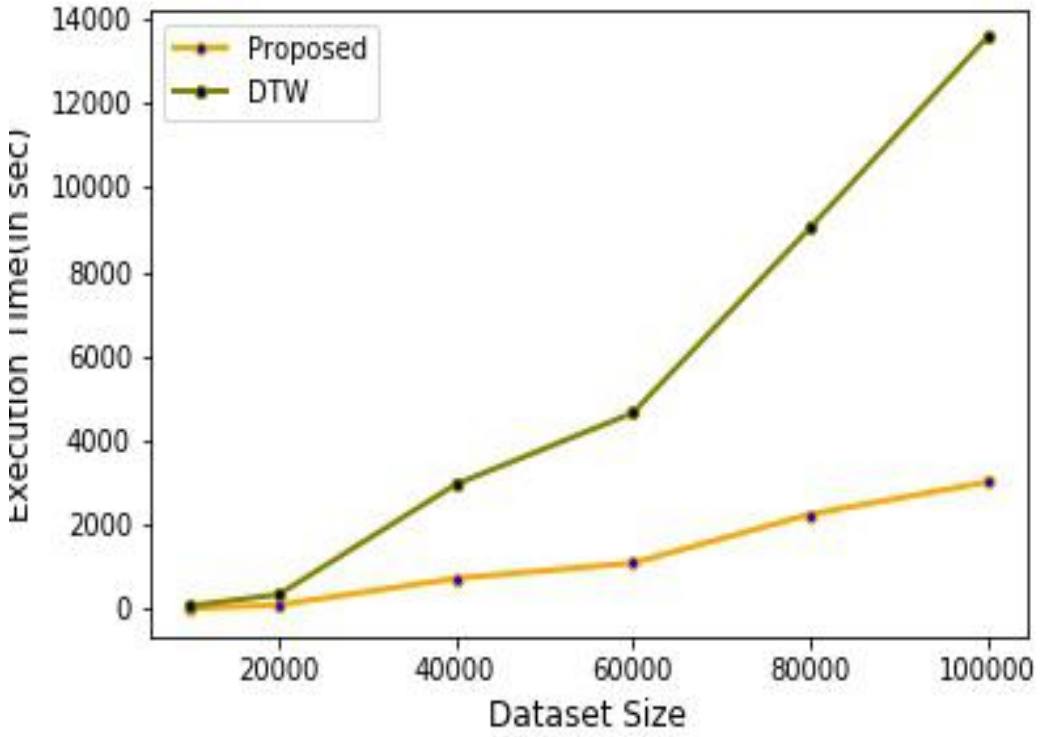
- Proposed approach produces less comparisons compared to Exhaustive Search using Dynamic Time Warping.
- Comparisons are directly proportional to the execution time.
- Proposed approach achieves dimension reduction of about 40% on original data.



Evaluation: Performance Evaluation

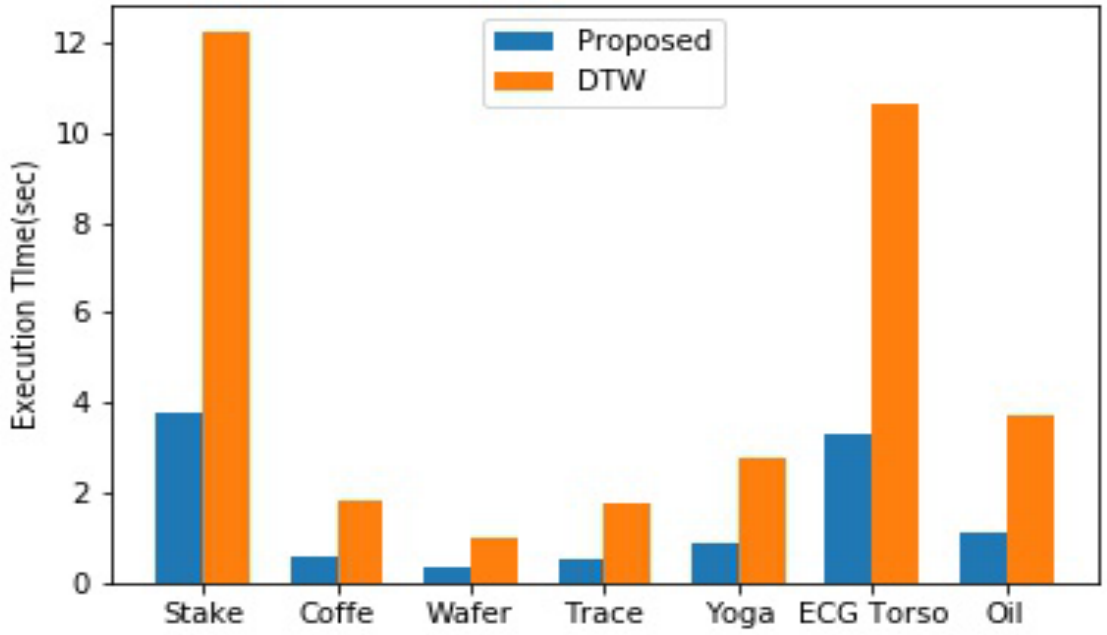
Execution time :

- As the dataset size increases, the execution time takes longer.
- Proposed approach has the lower computation time because of dimensionality reduction.
- Exhaustive Pattern Search by DTW takes 4 times higher execution time than the proposed approach.



Evaluation: Performance Evaluation

Computational speed on 7-UCR benchmark data sets :



Shortcomings:

- DTW is not a fastest, adding few more layers to improve it.
- Pattern detection in one dimensional time series data.

Conclusion

- Developed an approach to discover patterns of variable lengths and to prune non-overlapping patterns.
- Low computational time was achieved by symbolic representation using PAA.
- Patterns are ranked using DTW distance.
- To meet the user expectations and evaluate relevant patterns Feature vector is used for classification of the patterns.
- Experimental findings on data sets demonstrate that in contrast to state-of-the-art approach, our proposed approach finds out more significant patterns and rank them efficiently.

References

- Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. **Querying and mining of time series data: experimental comparison of representations and distance measures**. *Proceedings of theVLDB Endowment*, 1(2):1542–1552, 2008.
- Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. **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.*
- Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. **Dimensionality reduction for fast similarity search in large time series databases**. *Knowledge and information Systems, 3(3):263–286, 2001.*
- Yeh, Chin-Chia Michael, et al. **Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets**. *2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.*
- Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu,Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. ***The UCR time series archive***.

Thank you

Fraunhofer FIT