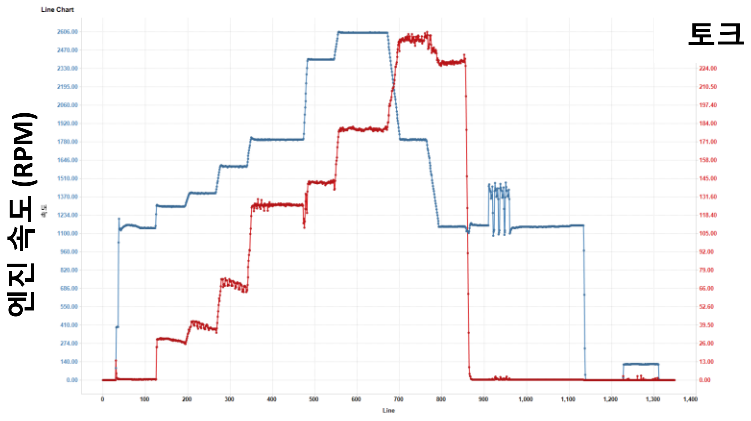
Multivariate time series classification

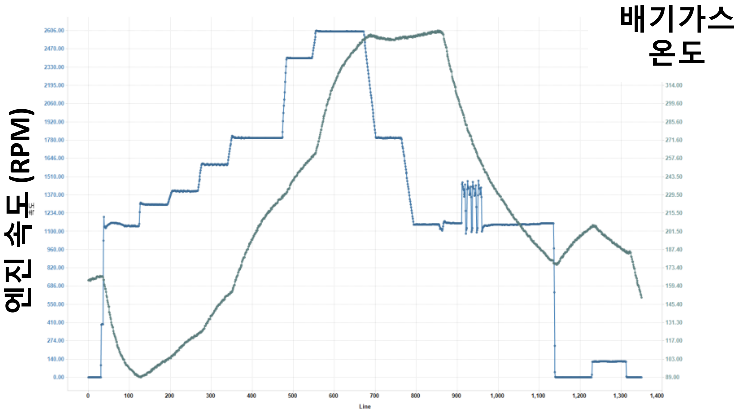
Oct 12, 2015

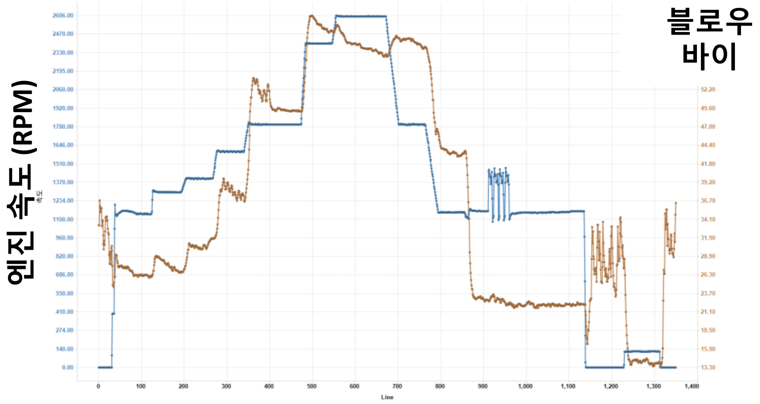
Je Hyuk Lee
Dept of Industrial Engineering, SNU

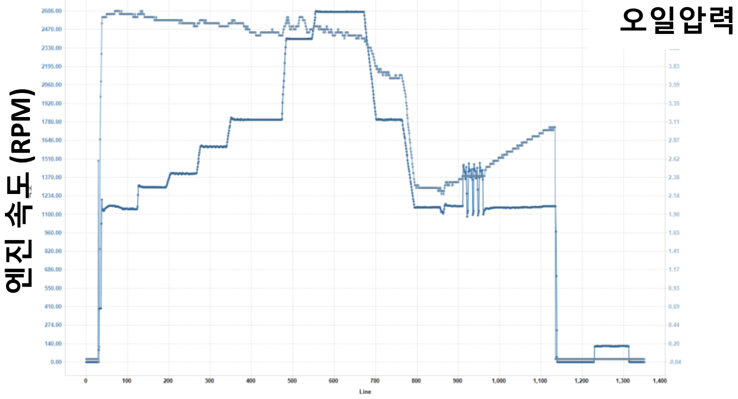
Data

- **Characteristics**
  - 엔진 속도를 step으로 변화시키면서, 여러 항목이 어떻게 변화하는지 측정
  - Multivariate Time series data









Questions

- **1. How to deal with multivariate time series data?**
  - Before the anomaly detection problem, how to solve classification problem?
  - classify multivariate time series
  - Spiegel, Stephan et al.(2011): classify multivariate time series

- **2. How to apply shapelet method to multivariate time series data?**
  - Ghalwash et al.(2012): multi-shapelet in bioinformatics
  - But what if the important feature is not obvious?

- **3. What if the important feature is not in each step?**
  - Only extracted features for each step
  - In the gap between two steps
  - Spiegel, Stephan et al.(2011): Segmentation and clustering using SVD

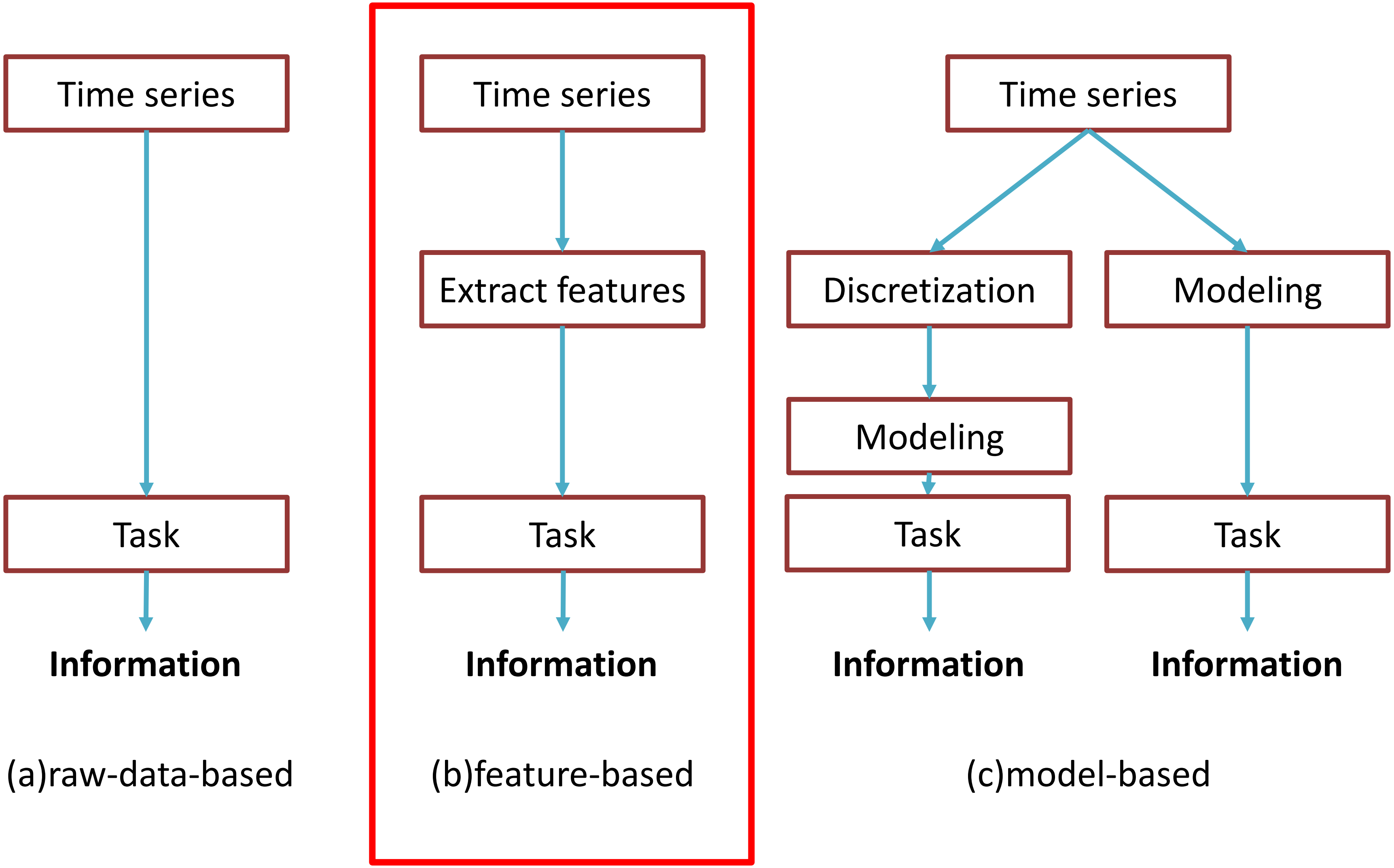
- **4. It has various different length. How to deal with this problem?**
  - …

Section 1

EARLY CLASSIFICATION OF MULTIVARIATE TIME SERIES BY SHAPELET

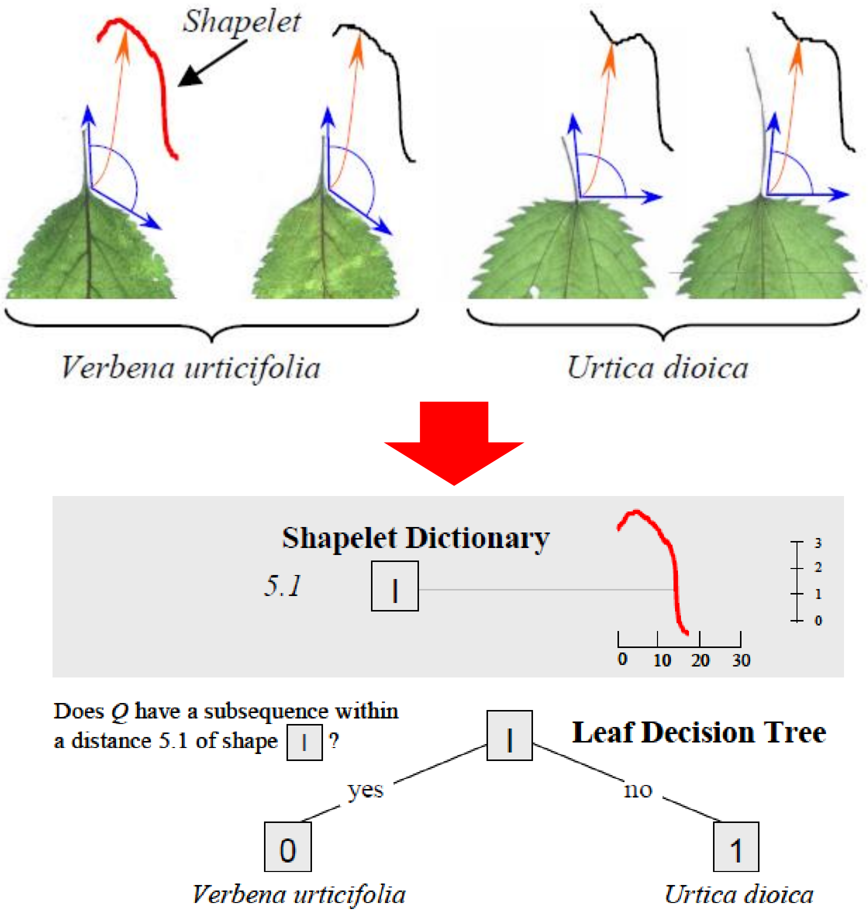
Ghalwash, Mohamed F., and Zoran Obradovic. "Early classification of multivariate temporal observations by extraction of interpretable shapelets."*BMC bioinformatics* 13.1 (2012): 195.

Time series data mining



Shapelet

- **Shapelet**
  - Subsequences which are maximally representative of a class
  - Motif of sequences



Does Q have a subsequence within a distance 5.1 of shape ?

Research

- Shapelet concept had been studied only in univariate study

- In this study,
  - Extend the concept of shapelet to multivariate case
  - Information-gain based distance threshold
  - Weighted information-gain based utility score of a shapelet

Univariate Shapelet

- **Formal definition**
  - T: Sequence data
  - S_p^l: subsequence that has starting point p, length l
  - $\boldsymbol{S_T^l}$: subsequences set. $\mathbf{S}_T^{\,l} = \{S_p^{\,l} \text{ of } T, \text{ for } 1 \leq p \leq m - l + 1\}$
  - Subsequence distance ($Subdist(T, S)$)
    - $\boldsymbol{Subdist(T,S) = \min\big(\mathbf{Dist}(\mathbf{S},\mathbf{S}')\big), \mathbf{where}\ \mathbf{S}' \in S_T^{|S|}}$
    - Minimum distance between S and subsequences of T which has length $|S|$
  - Entropy $I(\boldsymbol{D})$
    - $I(\boldsymbol{D}) = -p(A)\log\big(p(A)\big) - p(B)\log(p(B))$
  - Information Gain
    - Entropy difference when $\boldsymbol{D}$ is split into $\boldsymbol{D_1}, \boldsymbol{D_2}$ by split strategy sp
    - $\boldsymbol{Gain(sp) = I(D) - (f(D_1)I(D_1) + f(D_2)I(D_2))}$

Univariate Shapelet

- **Formal definition**
  - Shapelet is a kind of motif
    - Subsequence distance plays a important role in similarity measure
    - **How to set a starting point and length?**
  - How to classify Time series data
    - Data set $\boldsymbol{D}$는 class가 $\boldsymbol{A}$와 $\boldsymbol{B}$인 data point들로 구성됨
    - Classification Rule : $class = \begin{cases} \boldsymbol{D_1}, subsequenceDist(T_{1,i}, S) < d_{th} \\ \boldsymbol{D_2}, subsequenceDist(T_{1,i}, S) \geq d_{th} \end{cases}$
    - 분류된 class와 실제 class가 비슷한 distance threshold d_{th}를 찾아야
  - **Optimal split point** $(OSP(\boldsymbol{D}, S))$
    - Time series data set $\boldsymbol{D}$가 class $\boldsymbol{A}, \boldsymbol{B}$들로 이루어져 있다고 하자
    - A Shapelet candidate S에 대해서, 가장 분류를 잘하는 distance threshold
      - $\boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d}_{\boldsymbol{OSP}(\boldsymbol{D},\boldsymbol{S})}) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d}'_{\boldsymbol{th}})$, for any other distance threshold $\mathrm{d}'_{\mathrm{th}}$
  - **Shapelet** $(Shapelet(\boldsymbol{D}))$
    - 모든 candidate subsequence들과, 해당 OSP들 중, 가장 분류를 잘하는 subsequence
    - $\boldsymbol{Gain}\left(\boldsymbol{Shapelet}(\boldsymbol{D}), \boldsymbol{d}_{\boldsymbol{OSP}(\boldsymbol{D},\boldsymbol{Shapelet}(\boldsymbol{D}))}\right) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d}'_{\boldsymbol{th}})$

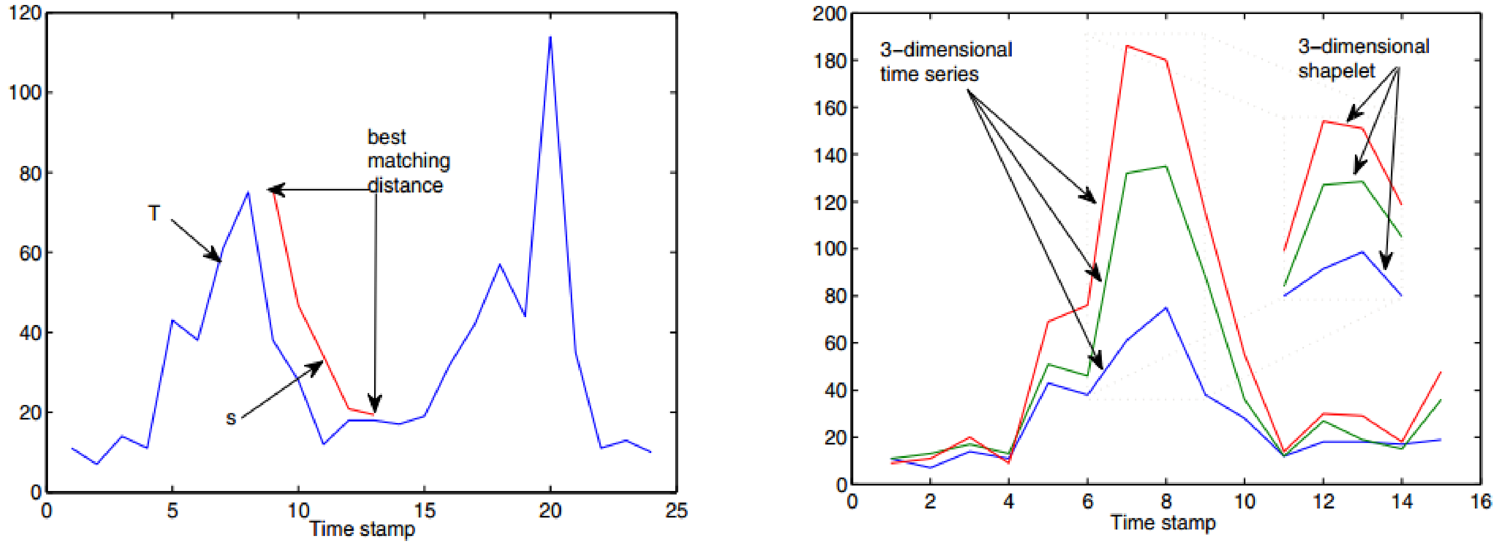
Multivariate Shapelet

- **Formal definition**
  - $T = [T^1, T^2, \ldots, T^N]$: N-dimensional sequence data
  - s^j: Subsequence의 j^{th} dimension factor
  - Subsequence distance $(Subdist(S, T))$
    - $\boldsymbol{Subdist(s, T)} = [\mathbf{dist}(\mathbf{s^1}, \mathbf{T^1}), \mathbf{dist}(\mathbf{s^2}, \mathbf{T^2}), \ldots, \mathbf{dist}(\mathbf{S^N}, \mathbf{T^N})]$
    - Minimum distance between S and subsequences of T which has length $|\boldsymbol{s^i}|$

Multivariate Shapelet

- **Formal definition**
  - Shapelet is a kind of motif
    - Subsequence distance plays a important role in similarity measure
  - How to classify Time series data
    - Data set $\boldsymbol{D}$는 class가 $\boldsymbol{A}$와 $\boldsymbol{B}$인 data point들로 구성됨
    - Classification Rule : $class = \begin{cases} \boldsymbol{D_1}, subsequenceDist\left(T_{1,i}, S\right) <_{perc} \boldsymbol{d_{th}} \\ \boldsymbol{D_2}, subsequenceDist\left(T_{1,i}, S\right) \geq_{perc} \boldsymbol{d_{th}} \end{cases}$
    - $d_1 <_{perc} d_2 \colon d_1^{qj} < d_2^{qj}, \forall j = 1 \ldots perc \times N, perc \in [0,1]$
    - 분류된 class와 실제 class가 비슷한 distance threshold $\boldsymbol{d_{th}}$를 찾아야
  - **Optimal split point** $(OSP(\boldsymbol{D}, S))$
    - Time series data set $\boldsymbol{D}$가 class $\boldsymbol{A}, \boldsymbol{B}$들로 이루어져 있다고 하자
    - A Shapelet candidate S에 대해서, 가장 분류를 잘하는 distance threshold
      - $\boldsymbol{Gain}\left(\boldsymbol{S}, \boldsymbol{d_{OSP(D,S)}}\right) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d'_{th}})$, for any other distance threshold $\mathrm{d'_{th}}$
  - **Shapelet** $(Shapelet(\boldsymbol{D}))$
    - 모든 candidate subsequence들과, 해당 OSP들 중, 가장 분류를 잘하는 subsequence
    - $\boldsymbol{Gain}\left(\boldsymbol{Shapelet(D)}, \boldsymbol{d_{OSP(D,Shapelet(D))}}\right) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d'_{th}})$
    - **여기서, 각 변수의 starting point는 같은 지점으로 함**

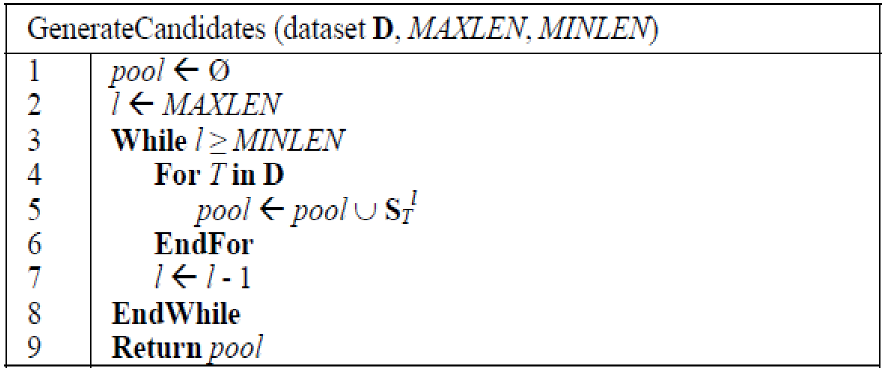
Example



(Left): Univariate case, (Right): Multivariate case

Shapelet - BruteForce Algorithm

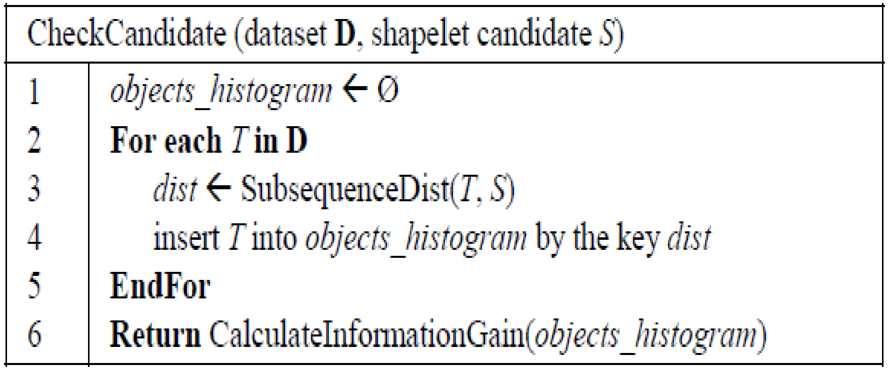
- **Generate Candidate**
  - 모든 shapelet 후보 생성



GenerateCandidates (dataset **D**, *MAXLEN*, *MINLEN*)

```
pool ← Ø
l ← MAXLEN
While l ≥ MINLEN
    For T in D
        pool ← pool ∪ $S_T^l$
    EndFor
    l ← l - 1
EndWhile
Return pool
```

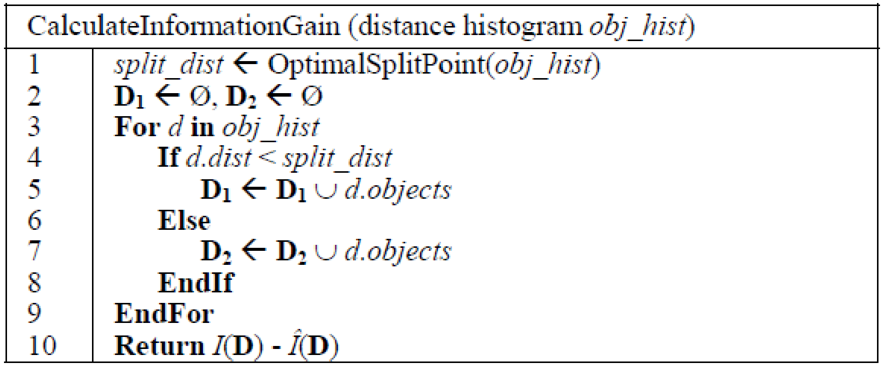
- **Check Candidate**
  - 후보 shapelet과의 dist계산
  - 이 dist들로 histogram 생성



CheckCandidate (dataset **D**, shapelet candidate *S*)

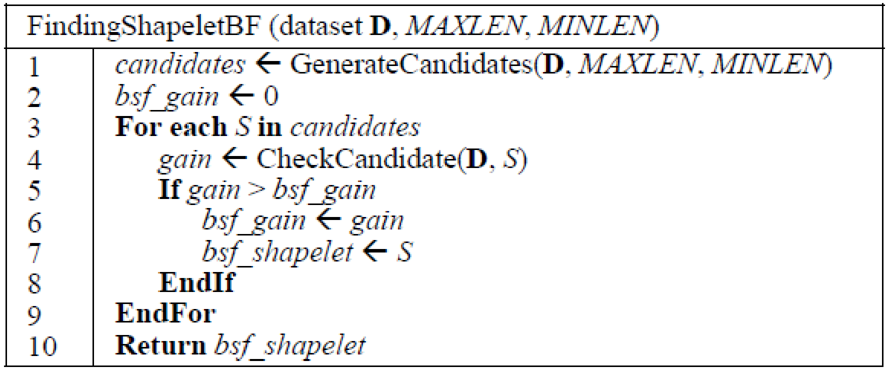
```
objects_histogram ← Ø
For each T in D
    dist ← SubsequenceDist(T, S)
    insert T into objects_histogram by the key dist
EndFor
Return CalculateInformationGain(objects_histogram)
```

Shapelet - BruteForce Algorithm



```
CalculateInformationGain (distance histogram obj_hist)
split_dist ← OptimalSplitPoint(obj_hist)
D₁ ← Ø, D₂ ← Ø
For d in obj_hist
    If d.dist < split_dist
        D₁ ← D₁ ∪ d.objects
    Else
        D₂ ← D₂ ∪ d.objects
    EndIf
EndFor
Return I(D) - Î(D)
```

- **CalculateInformationGain**
  - Split distribution set 생성
  - 각 split마다 분류
  - 각각 Information Gain 산출



```
FindingShapeletBF (dataset D, MAXLEN, MINLEN)
candidates ← GenerateCandidates(D, MAXLEN, MINLEN)
bsf_gain ← 0
For each S in candidates
    gain ← CheckCandidate(D, S)
    If gain > bsf_gain
        bsf_gain ← gain
        bsf_shapelet ← S
    EndIf
EndFor
Return bsf_shapelet
```

- **FindingShapeletBF**
  - 후보 shapelet 생성
  - 매 후보마다 IG 계산
  - IG, Shapelet 업데이트

Classification

- Score the shapelet candidates w.r.t. weighted information gain
  - Consider both early classification and accuracy

- If highest score can cover the current test time series
  - Classified as the class of the shapelet
  - Else, next highest score and repeats the process again

- (Maybe, it's not a good classification method)

Experiments

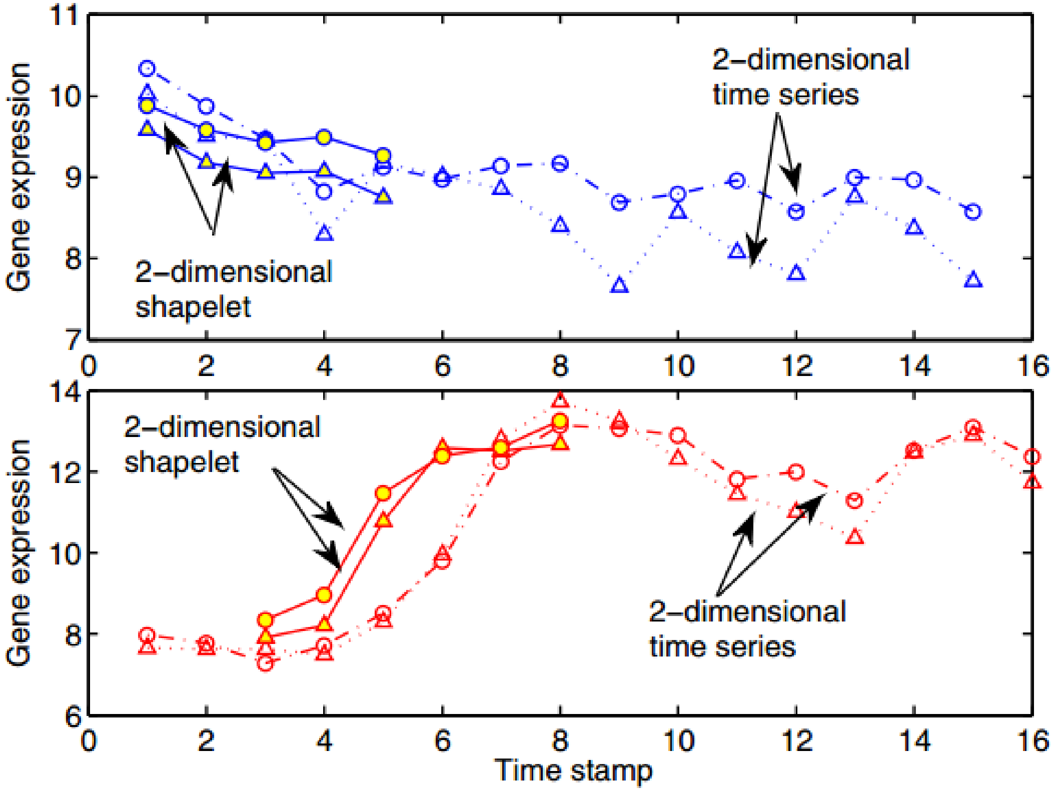
- 3 data sets
  - **Blood gene expression dataset** from human viral studies with influenza A (H3N2)
    - 17 subjects: 9 symptomatic, 8 asymptomatic
    - Blood samples taken from 16 time points
    - Used 23 unique genes from Zaas et al.(2009)
  - **HRV dataset** to distinguish acute respiratory infections
    - 20 subjects: 10 symptomatic, 10 asymptomatic
    - Blood samples taken from 14 time points
    - Used 26 unique genes from Zaas et al.(2009)
  - **Drug Response datasets** for drug therapy with IFNβ from Baranzini et al.(2005)
    - 52 patients: 33 good responders, 19 bad responders to the drug
    - Every 3 months in the first year, and every 6 months in the second year
    - Generated data (Baranzini 3A, 3B, 6, 12)
    - 9 genes that was founded from discriminative HMM study (Lin et al.(2008))
    - 17 relevant genes from mixture of HMM study (Costa et al.(2009))

Experiments

- Settings
  - Length : 3~60% of time series length
  - Bootstrapping for generalization error estimation
    - Sample with replacement from original dataset(75%)
    - Test with the others
    - Report the median of the accuracy
  - Report components
    - Accuracy
    - Coverage (percentage of the time series covered by the method)
    - Earliness (fraction of the time series length used for classification)
    - $F_1 = 2\frac{Acc(1-Ear)}{(1-Ear)+Acc}, 0 < F_1 < 1$

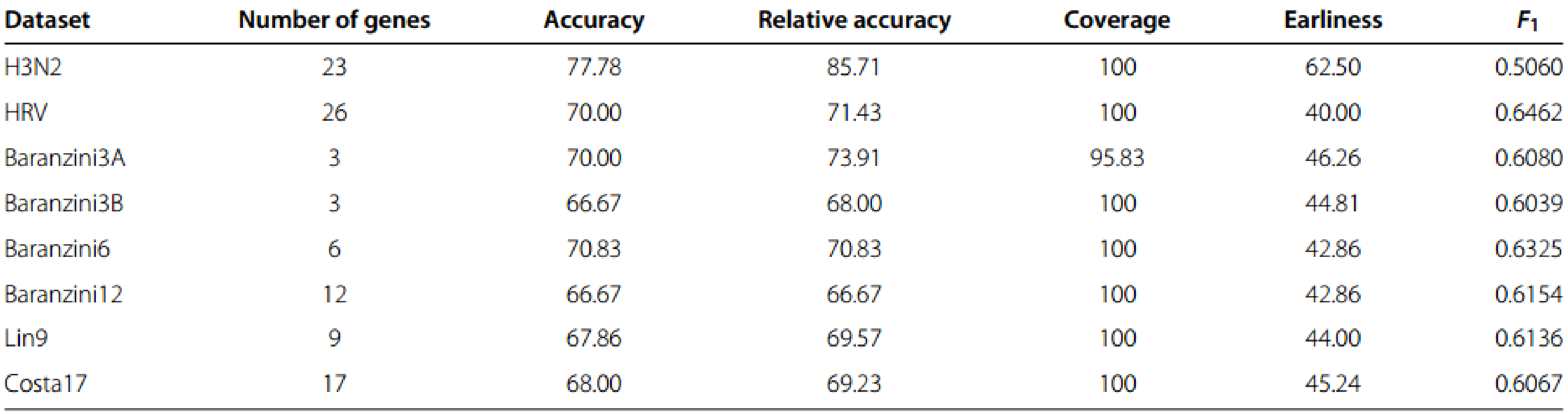
Experiments

- Results
  - Effectiveness of the MSD method on a case from H3N2 dataset
  - Used first half of the earliest time stamp
    - (Top): 2-D H3N2 asymptomatic test subject
    - (Bottom): 2-D H3N2 symptomatic test subject
    - Both used RSAD2 and IFI44L genes in each time step



Experiments

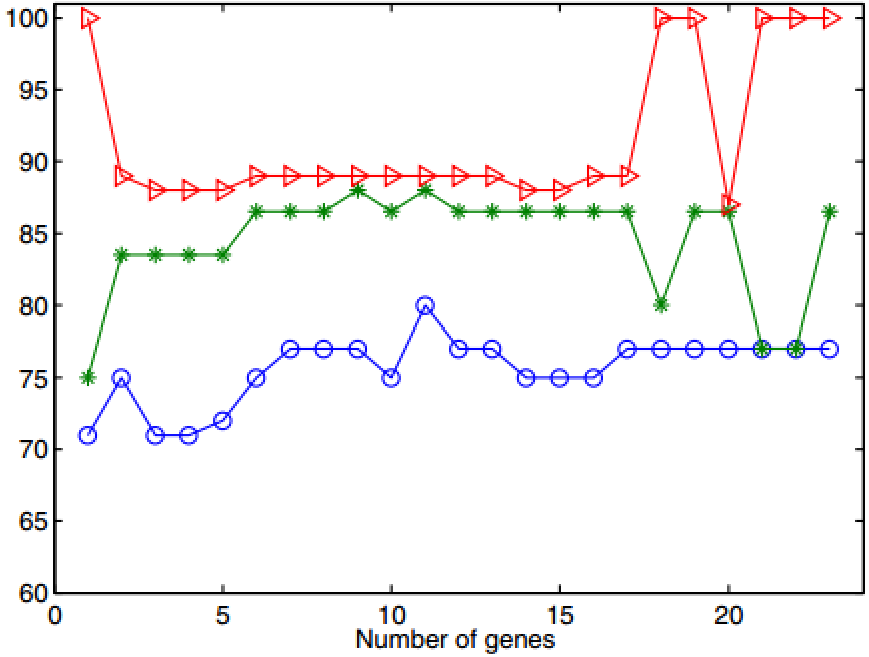
- Results
  - Performance of the MSD method on each dataset
    - Relatively good performance with a small fraction of the time series



Dataset	Number of genes	Accuracy	Relative accuracy	Coverage	Earliness	F_1
H3N2	23	77.78	85.71	100	62.50	0.5060
HRV	26	70.00	71.43	100	40.00	0.6462
Baranzini3A	3	70.00	73.91	95.83	46.26	0.6080
Baranzini3B	3	66.67	68.00	100	44.81	0.6039
Baranzini6	6	70.83	70.83	100	42.86	0.6325
Baranzini12	12	66.67	66.67	100	42.86	0.6154
Lin9	9	67.86	69.57	100	44.00	0.6136
Costa17	17	68.00	69.23	100	45.24	0.6067

Experiments

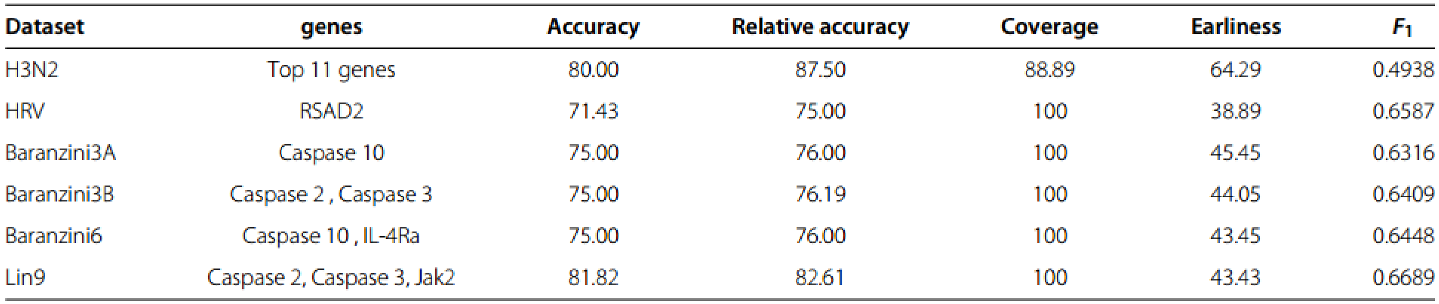
- Results
  - Performance of MSD on H3N2 dataset using top genes
    - Using some of the top genes rather than whole genes to get better performance
    - Top genes are studied by Zaas et al(2009)



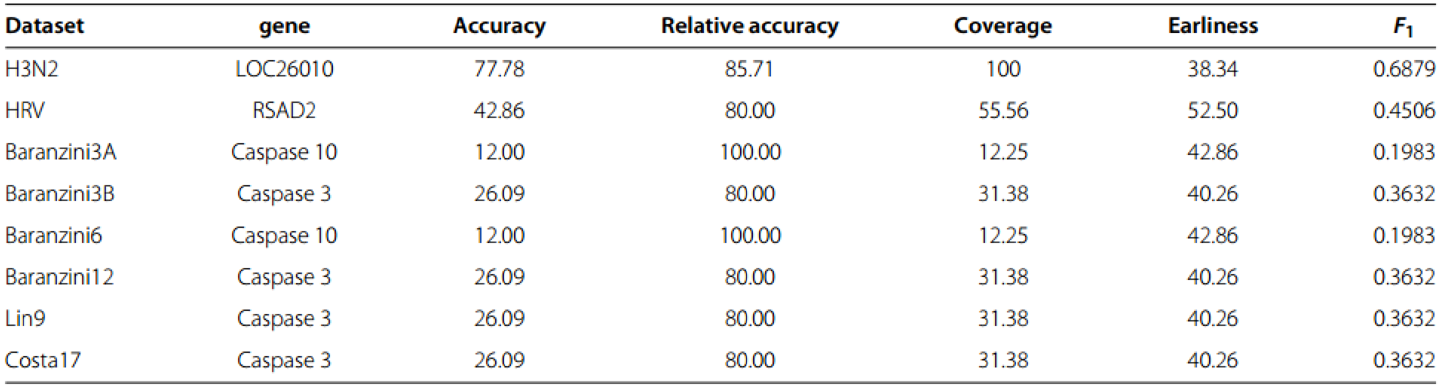
(Red): Coverage, (Green): Relative Accuracy, (Blue): Accuracy

Experiments

- Results
  - Comparing the MSD method with the univariate method
    - Without H3N2, MSD outperformed the univariate method
    - (Top): MSD method
    - (Bottom): Univariate method



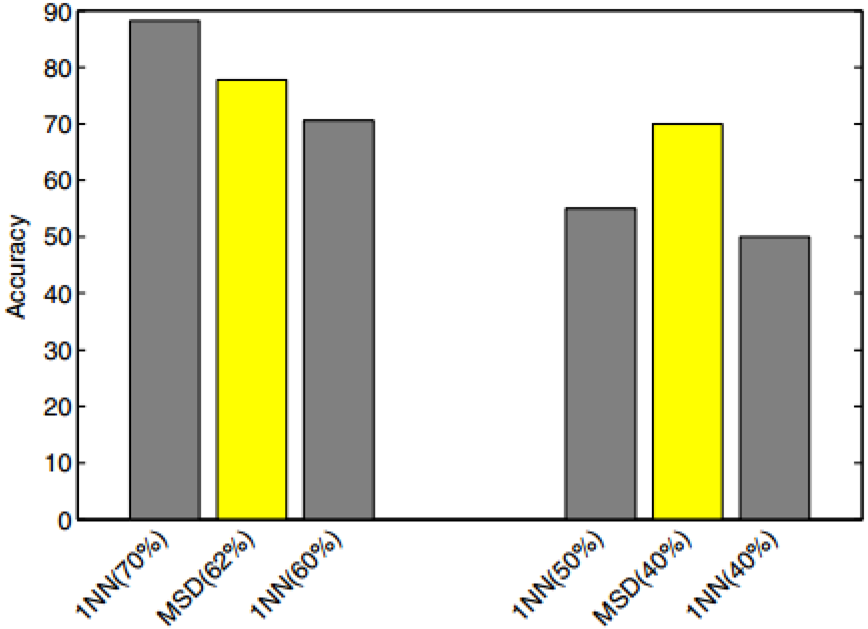
Dataset	genes	Accuracy	Relative accuracy	Coverage	Earliness	F_1
H3N2	Top 11 genes	80.00	87.50	88.89	64.29	0.4938
HRV	RSAD2	71.43	75.00	100	38.89	0.6587
Baranzini3A	Caspase 10	75.00	76.00	100	45.45	0.6316
Baranzini3B	Caspase 2 , Caspase 3	75.00	76.19	100	44.05	0.6409
Baranzini6	Caspase 10 , IL-4Ra	75.00	76.00	100	43.45	0.6448
Lin9	Caspase 2, Caspase 3, Jak2	81.82	82.61	100	43.43	0.6689



Dataset	gene	Accuracy	Relative accuracy	Coverage	Earliness	F_1
H3N2	LOC26010	77.78	85.71	100	38.34	0.6879
HRV	RSAD2	42.86	80.00	55.56	52.50	0.4506
Baranzini3A	Caspase 10	12.00	100.00	12.25	42.86	0.1983
Baranzini3B	Caspase 3	26.09	80.00	31.38	40.26	0.3632
Baranzini6	Caspase 10	12.00	100.00	12.25	42.86	0.1983
Baranzini12	Caspase 3	26.09	80.00	31.38	40.26	0.3632
Lin9	Caspase 3	26.09	80.00	31.38	40.26	0.3632
Costa17	Caspase 3	26.09	80.00	31.38	40.26	0.3632

Experiments

- Results
  - Comparing to conventional method
    - Comparing 1NN method with DTW which is exceptionally difficult to beat
    - To consider the earliness, used shorter time series
    - For the early classification task, MSD can be better than 1NN-DTW



Conclusion

- Shapelet method in multivariate time series is done
  - With simple concept
- This method is useful in early detection task
- Too much computation cost
  - Bruteforce approach: $O(k^2\overline{m}^3), \mathrm{k: number\ of\ subsequence}, \overline{m}: average\ length$
  - Only can be applied in short time series
  - Need more efficient algorithm

To dos

- Only for independence among attributes
- Too much computational cost for original task
- Difference in time series length
- How can DTW replace the Euclidean distance?
- Focusing on small data set with classification task

References

- Baranzini, Sergio E., et al. "Transcription-based prediction of response to IFNβ using supervised computational methods." (2004): e2.
- Costa, Ivan G., et al. "Constrained mixture estimation for analysis and robust classification of clinical time series." *Bioinformatics* 25.12 (2009): i6-i14.
- Ghalwash, Mohamed F., and Zoran Obradovic. "Early classification of multivariate temporal observations by extraction of interpretable shapelets."*BMC bioinformatics* 13.1 (2012): 195.
- Lin, Tien-ho, Naftali Kaminski, and Ziv Bar-Joseph. "Alignment and classification of time series gene expression in clinical studies." *Bioinformatics*24.13 (2008): i147-i155.
- Ye, Lexiang, and Eamonn Keogh. "Time series shapelets: a novel technique that allows accurate, interpretable and fast classification." *Data mining and knowledge discovery* 22.1-2 (2011): 149-182.
- Zaas, Aimee K., et al. "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans." *Cell host & microbe*6.3 (2009): 207-217.