Research Proposal

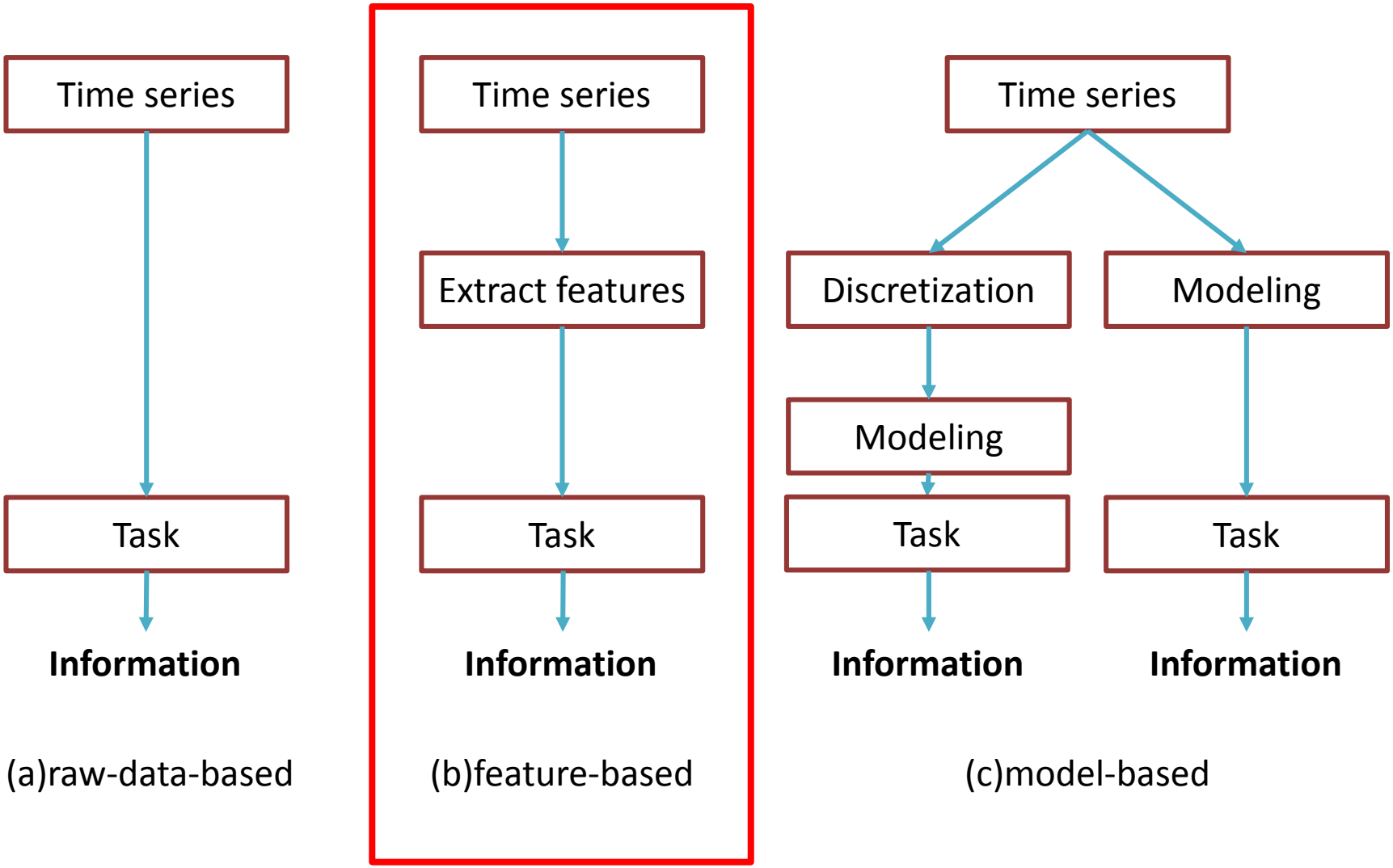
Sep 14, 2015

Je Hyuk Lee
Industrial Engineering Department, SNU

Section1

TIME SERIES MOTIFS FOR QUALITY IMPROVEMENT

Time series data mining

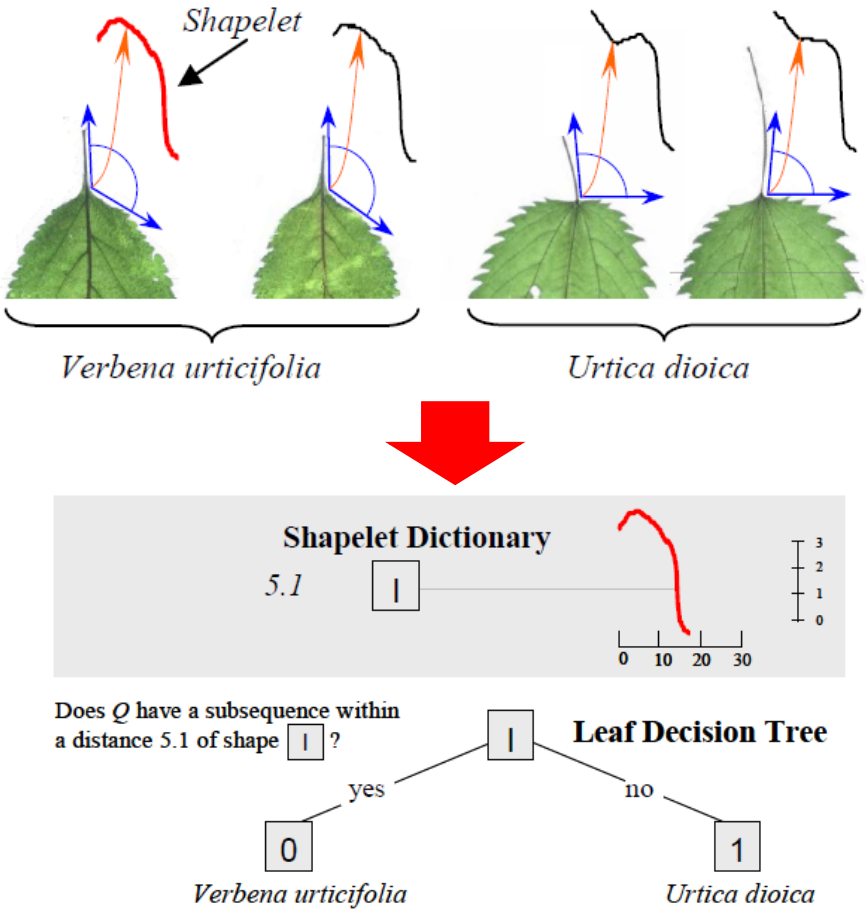


Feature-based method

- **Several Feature-based method**
  - Statistics (mean, std, max, min…)
  - Discrete Fourier Transform (DFT)
  - Discrete Wavelet Transform (DWT)
  - SAX
  - …
- **Similarity measure**
  - Euclidean Distance
  - Dynamic Time Warping (DTW)
  - …
- **Simple nearest neighbor algorithm is the most accurate method**
  - (Keogh et al., 2008)
  - But, too much cost
  - Also, does not significantly outperform random guessing
  - Because of noise!

Shapelet

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



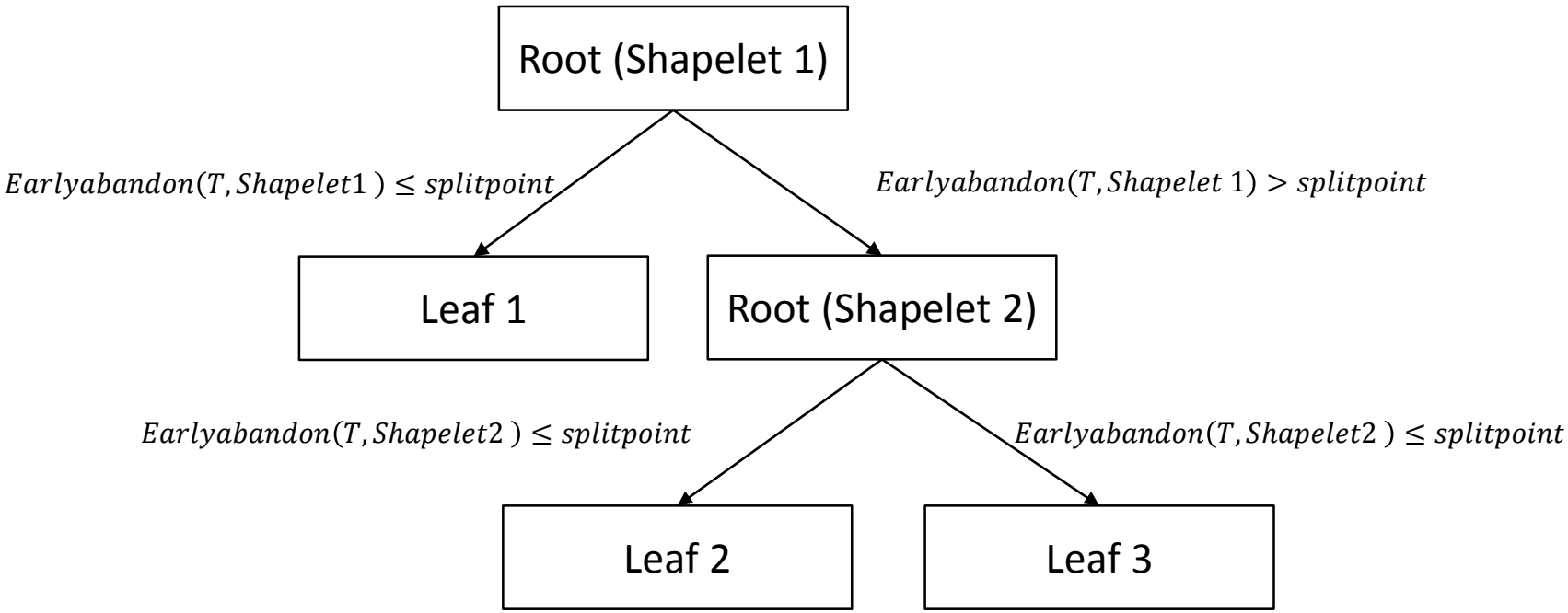
Shapelet

- **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
      - $\boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d_{OSP(D,S)}}) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d'_{th}})$, for any other distance threshold $\mathrm{d'_{th}}$
  - **Shapelet** $(Shapelet(\boldsymbol{D}))$
    - 모든 candidate subsequence들과, 해당 OSP들 중, 가장 분류를 잘하는 subsequence
    - $\boldsymbol{Gain}\left(\boldsymbol{Shapelet(D)}, \boldsymbol{d_{OSP(D,Shapelet(D))}}\right) \geq \boldsymbol{Gain}(\boldsymbol{S}, \boldsymbol{d'_{th}})$

Classification

- **분류 방법론: Decision Tree**
  - 각 DT step마다 shapelet을 생성하여 training
  - 기존에 존재하는 DT방법론 (Geurts and Pierre, 2001)

- **Prediction**



To-dos

- Shapelet algorithm구현
  - DP?
- Online Shapelet
  - 지속적인 업데이트가 필요
  - E.Keogh의 online shapelet 논문
- Data library에 적용
- 실제 anomaly detection문제에 적용하여, 기존에 했던 방법론과 비교