nTreeClus: A Model-based clustering of sequential data

Hadi Jahanshahi

Supervisor: Mustafa Gökçe Baydoğan

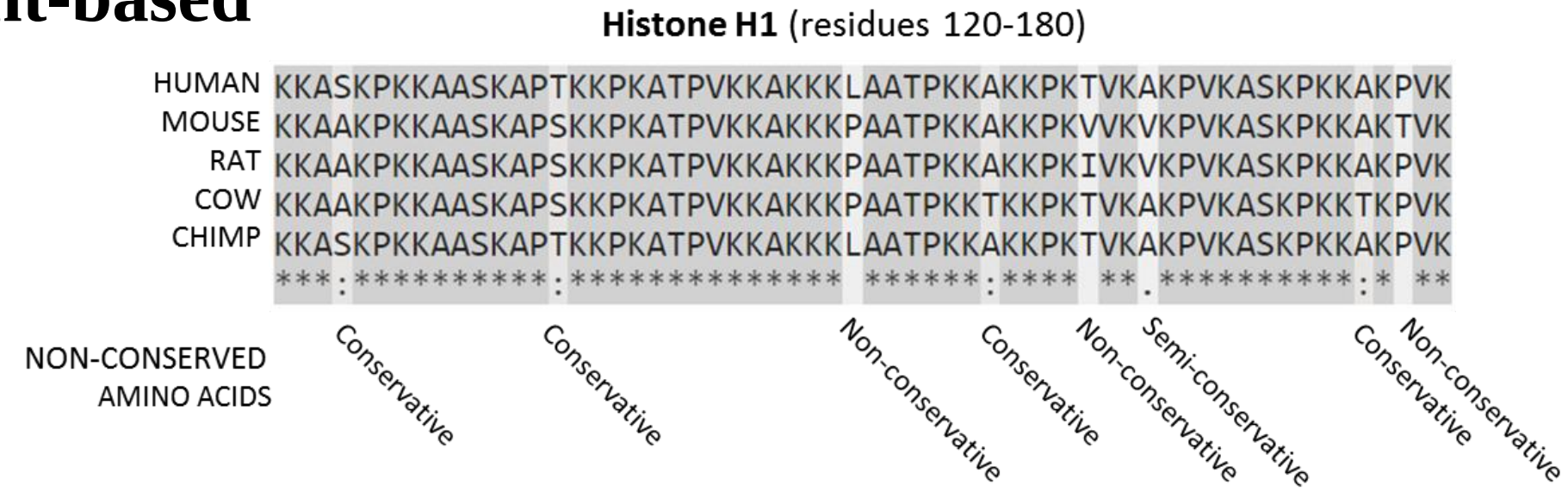
January 2020

Sequential Data

- Why sequence mining is important?
  - They are omnipresent:
    - DNA, Protein, purchase history, click-stream, flows of data, etc.
  - There are ways to improve current methods.
- Definition
  - A sequence over an element type A is an ordered list $X = x_1 \ldots x_m$, where
    - each x_1 is a member of $A=\{a_1,\ldots,a_n\}$, and is called an element of X
    - m referred to as the length of X.
- Types of methods for sequence mining/clustering
  - Model-based approaches
  - Proximity-based approaches:
    - Similarity-based and Feature-based

Similarity-based (I)

1. Alignment-based



- Limiting factors:
  - Generality issue, number of sequences, computation time, position of the pattern, and length of sequences.

Similarity-based (II)

2. Distance metrics:

Hamming distance

ATCGGTAGT
ATGGTTCCT

Limiting factors: only substitution, sequence of the same length

Levenshtein Distance

INTENTION → INTEN TION
EXECUTION → EXECUTION

Limiting factors: computational complexity, median string is not tractable

Jaro-Winkler

$$d_j = \frac{1}{3}\left(\frac{m}{|string_1|} + \frac{m}{|string_2|} + \frac{m-t}{m}\right)$$

$$d_{jw} = d_j + (lp(1-d_j))$$

m: matching characters
t: the number of transpositions
l: the length of common prefix
p: constant scaling factor

Limiting factors: good for short text, fails in many real cases

Feature-based

1. k-mers (n-gram)

GAATTCTCTGTAACCTCAGAGGTAGATAGA

GAA	3-mer	AAA	3-spectrum	0
AAT		AAG		0
ATT		AAT		0
TTC		AAC	3^4 combinations	1/81
TCT		....		...
CTC		CCA		0
TCT		CCG		0
CTG		CCT		1/81
28 states	....			

2. k-gapped pair

$$F_{i,j}^{k} = \frac{1}{\mathcal{M}-k-1} \sum_{m=1}^{\mathcal{M}-k-1} O_{i,j}(m, m+k+1)$$

$O_{i,j}(m, m+k+1)$: 1 if $x^m = i$ & $x^{m+k+1} = j$; 0 otherwise

$\frac{1}{\mathcal{M}-k-1}$: the length of sequence

Limiting factors: lower k and higher k issues:

Space required to store DNA, so much or so few overlapping

Complexity increases as the number of alphabets increases.

Autoregressive models

- AR models indicates that the output variable depends on its own previous values and because of its stochastic nature, it takes advantage of the statistics of the data.

$$X_{t+1} = b_0 + b_1 X_{t-1} + b_2 X_{t-2}$$

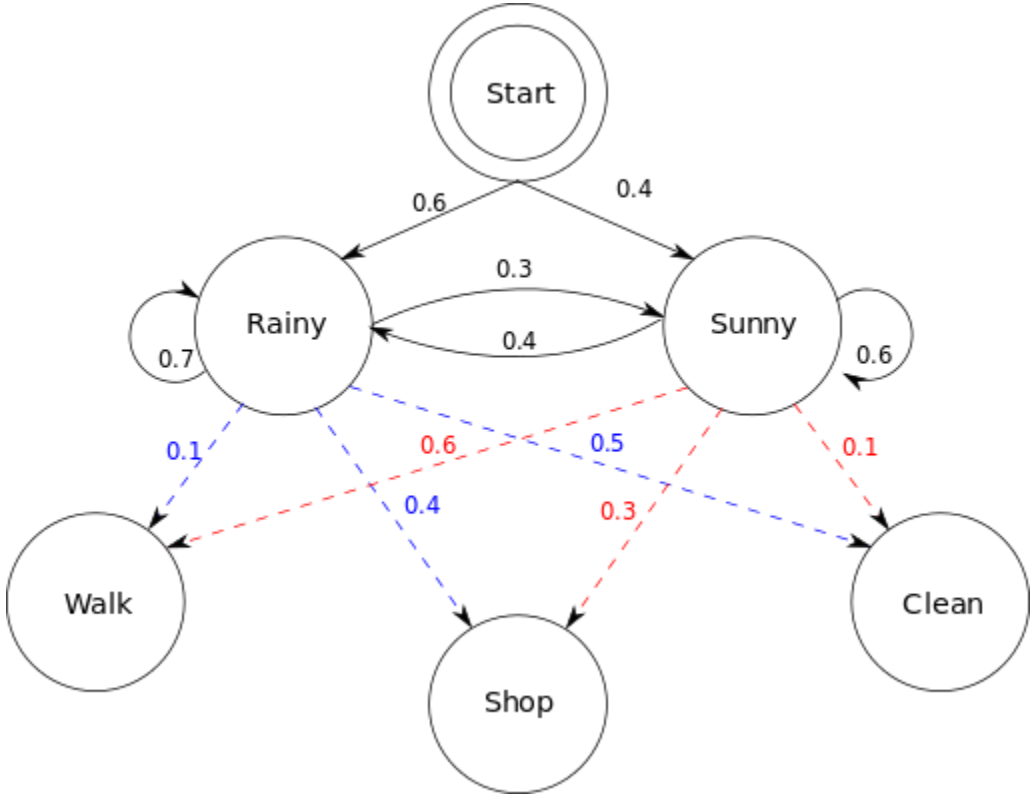
- It includes two steps:
  - Numerical mapping
  - AR model implementation

$$A = 1,\ G = 2,\ C = 3,\ and\ T = 4$$
$$A = -1.5,\ G = 1.5,\ C = 0.5,\ and\ T = -0.5$$

Limiting factors: irrational deduction based on numerical / binary mapping; setting model order can be problematic

Mixture HMM

- Initial probs
- Transition probs
- Emission probs



Limiting factors: computationally intensive, hard to be tuned

nTreeClus – defining sequence

$$X_{\mathcal{L}}^{\mathcal{M}} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & \cdots & x_1^{m-1} & x_1^m \\ x_2^1 & x_2^2 & \cdots & \cdots & x_2^{m-1} & x_2^m \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ x_{l-1}^1 & x_{l-1}^2 & \cdots & \cdots & x_{l-1}^{m-1} & x_{l-1}^m \\ x_l^1 & x_l^2 & \cdots & \cdots & x_l^{m-1} & x_l^m \end{bmatrix}$$

- An example:

  evidence, evident, provide, unconventional and convene

$$\boldsymbol{X_5^{14}} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{N} & \boldsymbol{C} & \boldsymbol{O} & \boldsymbol{N} & \boldsymbol{V} & \boldsymbol{E} & \boldsymbol{N} & \boldsymbol{T} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{N} & \boldsymbol{A} & \boldsymbol{L} \\ \boldsymbol{C} & \boldsymbol{O} & \boldsymbol{N} & \boldsymbol{V} & \boldsymbol{E} & \boldsymbol{N} & \boldsymbol{E} & & & & & & & \\ \boldsymbol{P} & \boldsymbol{R} & \boldsymbol{O} & \boldsymbol{V} & \boldsymbol{I} & \boldsymbol{D} & \boldsymbol{E} & & & & & & & \\ \boldsymbol{E} & \boldsymbol{V} & \boldsymbol{I} & \boldsymbol{D} & \boldsymbol{E} & \boldsymbol{N} & \boldsymbol{C} & \boldsymbol{E} & & & & & & \\ \boldsymbol{E} & \boldsymbol{V} & \boldsymbol{I} & \boldsymbol{D} & \boldsymbol{E} & \boldsymbol{N} & \boldsymbol{T} & & & & & & & \end{bmatrix}$$

$$\mathrm{M} = max\{15, 7, 7, 8, 7\} = 15$$

$$A = \{u, n, c, o, n, v, e, t, i, a, l, p, r, d\} \Rightarrow a = 14$$

nTreeClus – Segmentation (I)

$$X_1 : abcaaabcbaa \xrightarrow{\text{windows size} = 5}$$

1^*	a	b	c	a	a	a
2^*	b	c	a	a	a	b
3^*	c	a	a	a	b	c
4^*	a	a	a	b	c	b
5^*	a	a	b	c	b	a
6^*	a	b	c	b	a	a

nTreeClus – Segmentation (II)

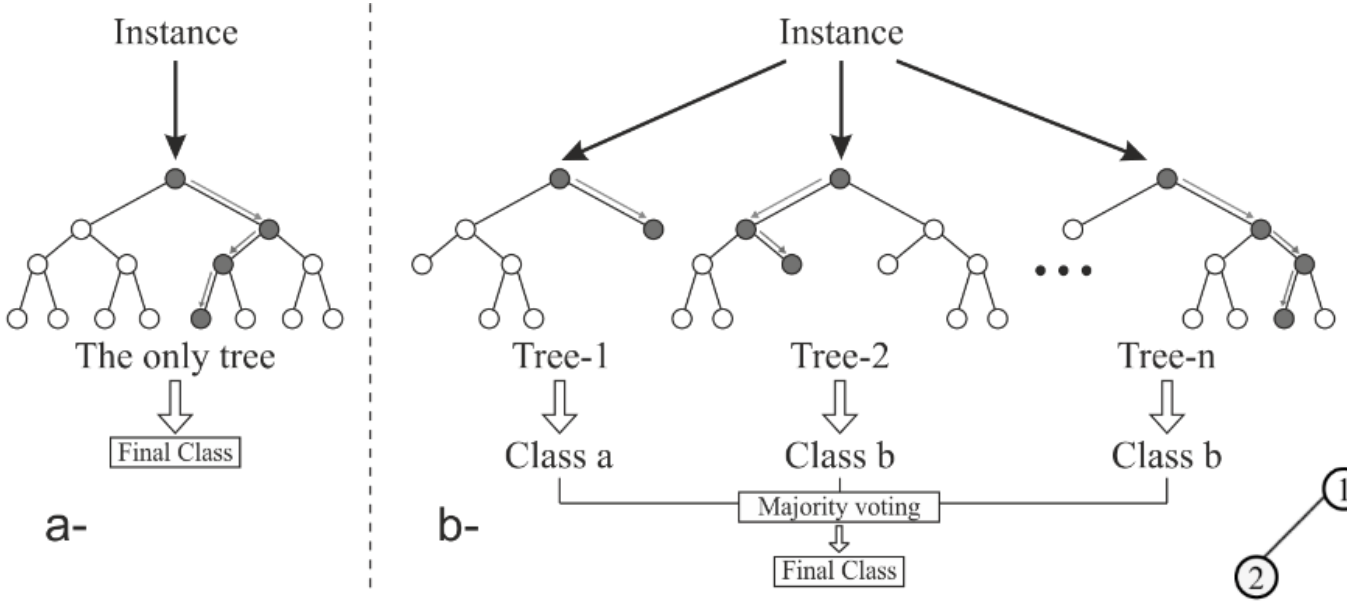
Algorithm 1 Matrix Segmentation

Data: sequential/categorical dataset (D) of size $\mathcal{L} \times \mathcal{M}$, the segmentation length (n) (*with default value of* $\sqrt{\mathcal{M}}$).
Result: Segmented matrix (length-smoothed sequences)

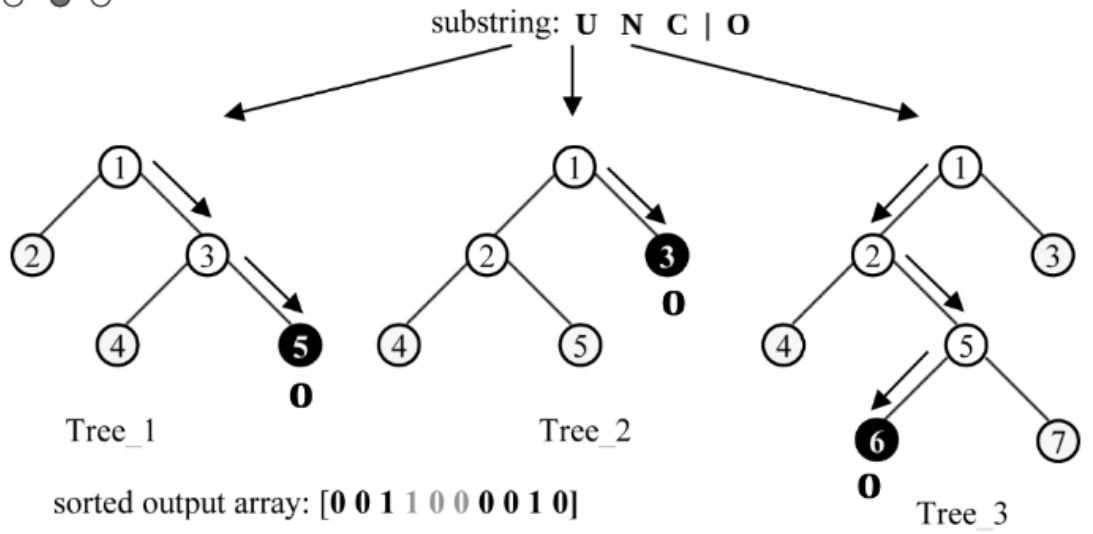
```
initialization;
  Segmented Matrix = Ø
for i ∈{1,..,L} do
    for j ∈{1,..,(M_i-n)} do
        Temporary Row ← Dataset[i,j:j+n]
        Temporary Row['y']← i
          if the position of element is important then
              Temporary Row['position']← j
          end if
        Append [Temporary Row] to Segmented Matrix
    end
end
```

nTreeClus – representation (I)

DT | RF



substring's id	tree 1	tree 2	tree 3
1 (abcaa)	5	3	6
2 (bcaaa)	2	5	6
3 (caaab)	5	5	6
4 (aaabc)	5	5	7
5 (aabcb)	2	3	7
6 (abcba)	4	3	6



nTreeClus – representation (II)

Algorithm 2 nTreeClus Representation

Data: Segmented Matrix of size $((\mathcal{L} \times (\mathcal{M} - n)) \times (n + 2))$, number of trees (t) *with default value of 10.*

Result: nTreeClus representation of dataset D (nTreeClus Representation)

1 initialization;

xtrain ← Segmented Matrix [:,1:n]

ytrain ← Segmented Matrix [:,n+1]

2 Train RandomForest (X=xtrain, Y=ytrain, ntree = t)

3 terminalRF ← trace the path of each row in Segmented Matrix to the terminal node for all t trees and store it.

4 nTreeClus Representation = An empty DataFrame whose number of columns is equal to terminalRF's one and whose number of rows is equal to Dataset's one.

5 **for** $i \in \{1,..,\mathcal{L}\}$ **do**

6 nTreeClus Representation [i,:] = $\sum_{j=1}^{(\mathcal{L}\times(\mathcal{M}-n))}$ terminalRF$[j,:]$ iff Segmented Matrix [j,'y']==i

7 **end**

$$\eta_{(abcaaabcbaa)} = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\downarrow \Sigma$$

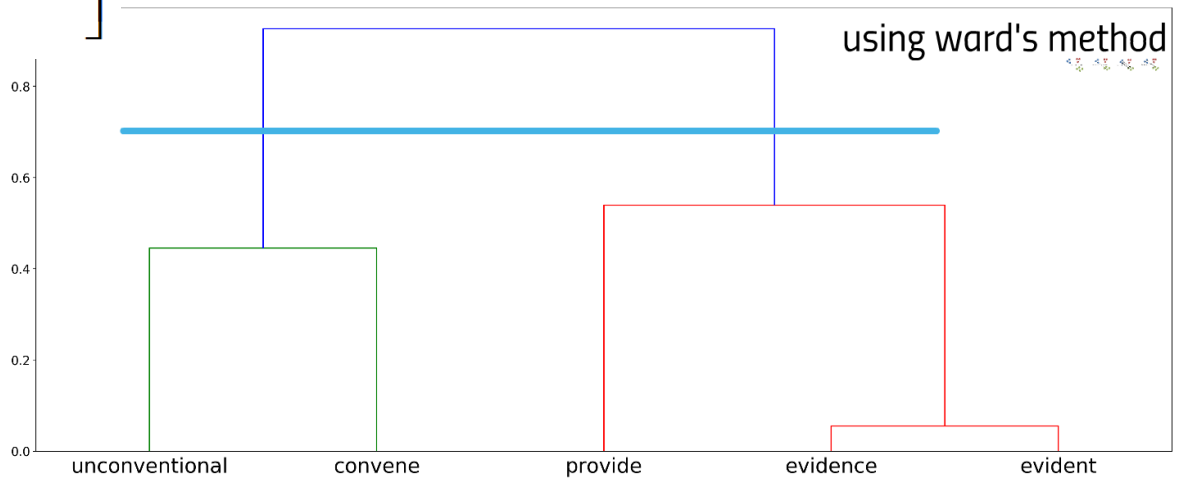
$$\eta_{aggregated} = \begin{pmatrix} 2 & 1 & 3 & 3 & 0 & 3 & 0 & 0 & 4 & 2. \end{pmatrix}$$

nTreeClus – Hierarchical Clustering

- After applying cosine similarity, hierarchical clustering has been conducted and the related dendrogram is shown below.

	evidence	*evident*	*provide*	*unconventional*	*convene*
evidence	**0**	**0.10101553**	**0.52973003**	**0.59819281**	**0.70042766**
evident		**0**	**0.51043498**	**0.6744861**	**0.73162748**
provide			**0**	**0.71894141**	**0.8179937**
unconventional				**0**	**0.27053775**
convene					**0**

$$cos(\theta) = 1 - \frac{a.b}{||a||_2||b||_2} = 1 - \frac{\sum_{i=1}^{l} a_i \times b_i}{\sqrt{\sum_{i=1}^{l}(a_i)^2} \times \sqrt{\sum_{i=1}^{l}(b_i)^2}}$$



Performance Metrics

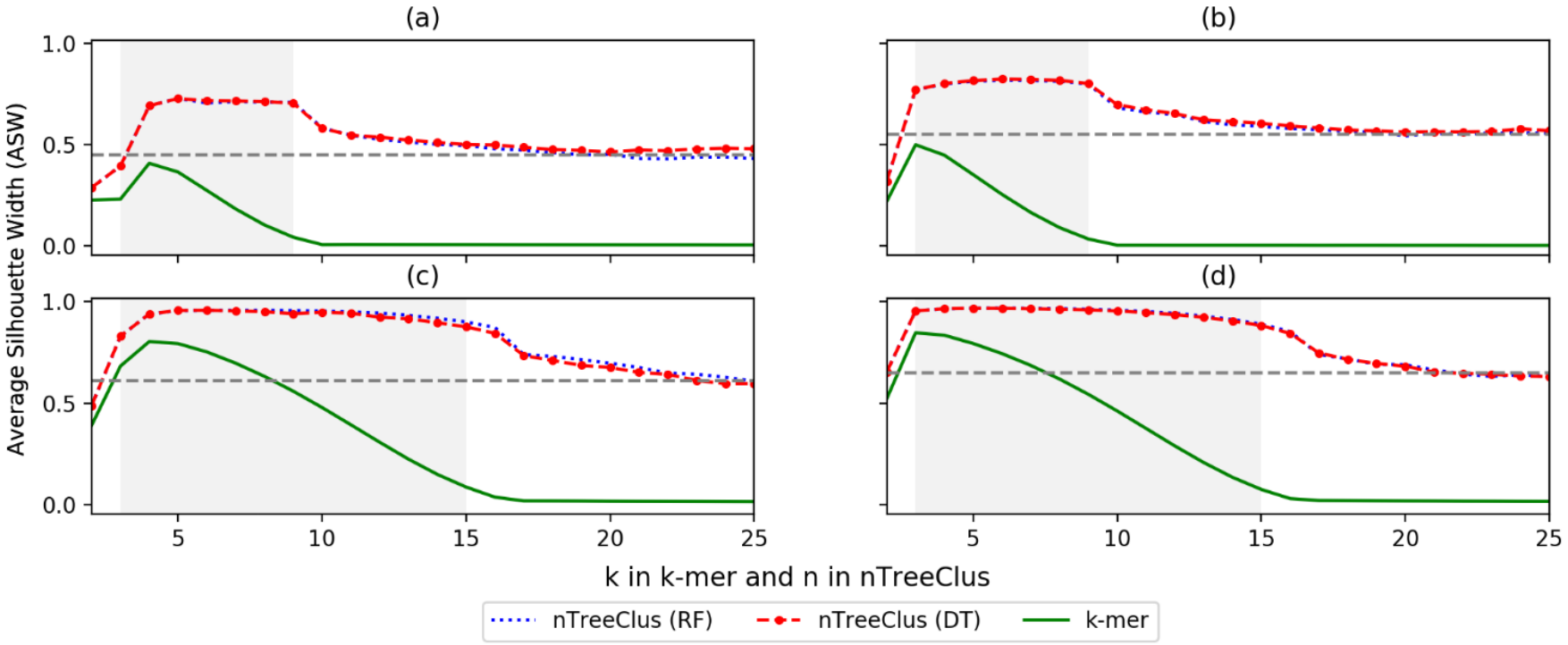
- Internal cluster validation indexes
  - Calinski-Harabasz Index (CH)
  - Average Silhouette Width (ASW)
  - Dunn Index (DI)

- External cluster validation indexes
  - Purity
  - Rand Index (RI)
  - Adjusted Rand Index (ARI)
  - F-measure

- Comparing with Nearest Neighbor Classifier

EXP I - Tuning Parameter n of nTreeClus

- Comparing the sensitivity of nTreeClus with parameter n and k-mer with parameter k



R = 10
N = 40
L = 180
M = 40
M_p = {8 , 15}
a = {7 , 20}
C = 2

R: the number of replications; N: number of batches; L: number of sequences in each batch;

M: Length of sequence; M_p: Length of pattern, a: length of alphabet set, C: number of clusters

EXP II – Pattern recognition

Methods	External Validation				Internal Validation	1NN
	Purity	RI	ARI	F-meas	ASW	
nTreeClus (DT)	0.988	0.980	0.959	0.987	**0.655**	0.986
nTreeClus (RF)	**0.992**	**0.984**	**0.969**	**0.990**	0.651	**0.989**
nTreeClus (DT)*	0.973	0.949	0.898	0.962	0.529	0.978
nTreeClus (RF)*	0.984	0.970	0.940	0.979	0.546	0.986
Levenshtein	0.892	0.830	0.660	0.868	0.343	0.932
Jaro-Winkler	0.654	0.497	-0.006	0.508	0.007	0.234
k-mers						
$k = 1$	0.793	0.688	0.377	0.762	0.385	0.740
$k = 2$	0.910	0.861	0.721	0.901	0.420	0.894
$k = 3$	0.967	0.945	0.891	0.963	0.442	0.958
$k = \sqrt{\mathcal{M}_i}$	0.971	0.936	0.873	0.939	0.346	0.937
MHMM	0.960	0.849	0.700	0.816	-	-

R = 10
N = 360
L = {40, 120, 200}
M = [80, 120]
M_p = {6, 10, 20}
α = {5, 7, 11, 16}
C = 2

EXP III – Gapped Pattern

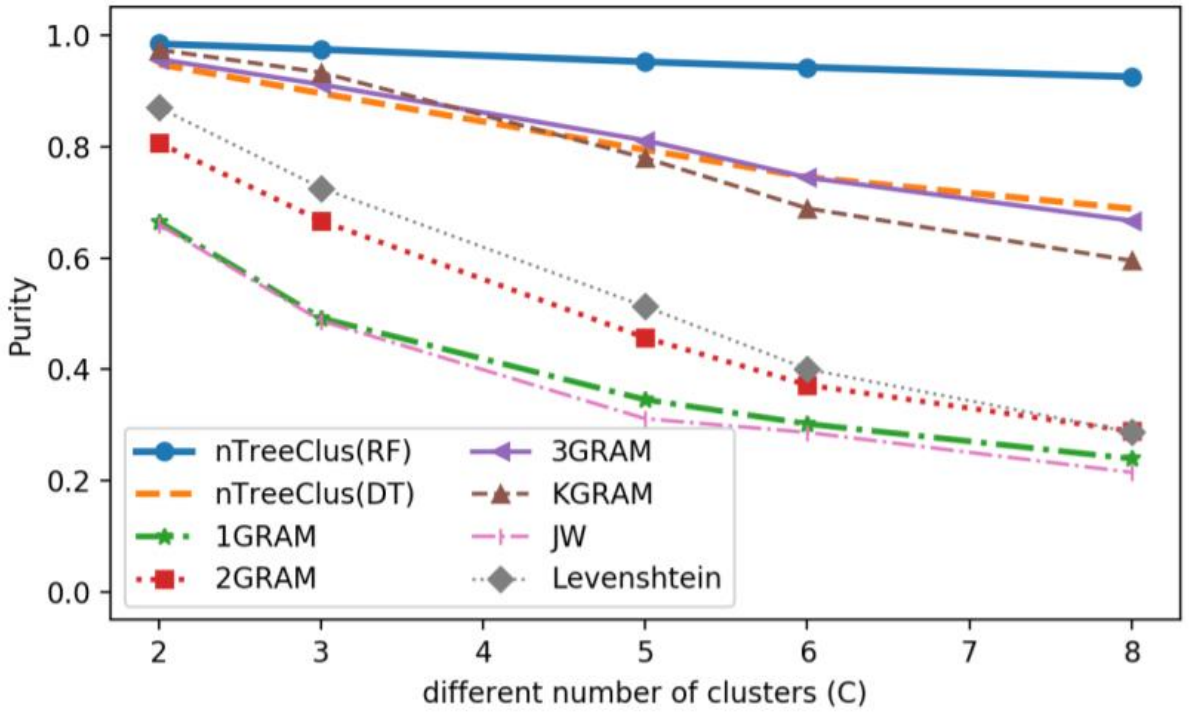
Methods	External Validation				Internal Validation	1NN
	Purity	RI	ARI	F-meas	ASW	
nTreeClus (DT)	0.939	0.885	0.771	0.904	**0.473**	0.909
nTreeClus (RF)	**0.943**	**0.899**	**0.798**	**0.916**	0.466	**0.920**
nTreeClus (DT)*	0.923	0.853	0.707	0.879	0.387	0.903
nTreeClus (RF)*	0.935	0.881	0.763	0.902	0.378	0.919
Levenshtein	0.869	0.796	0.592	0.839	0.319	0.868
Jaro-Winkler	0.655	0.496	-0.006	0.507	0.007	0.241
k-mers						
$k = 1$	0.788	0.677	0.355	0.748	0.374	0.728
$k = 2$	0.888	0.825	0.650	0.870	0.368	0.864
$k = 3$	0.927	0.886	0.773	**0.917**	0.335	0.907
$k = \sqrt{\mathcal{M}_i}$	0.902	0.774	0.549	0.777	0.195	0.808
MHMM	**0.943**	0.783	0.571	0.732	-	-

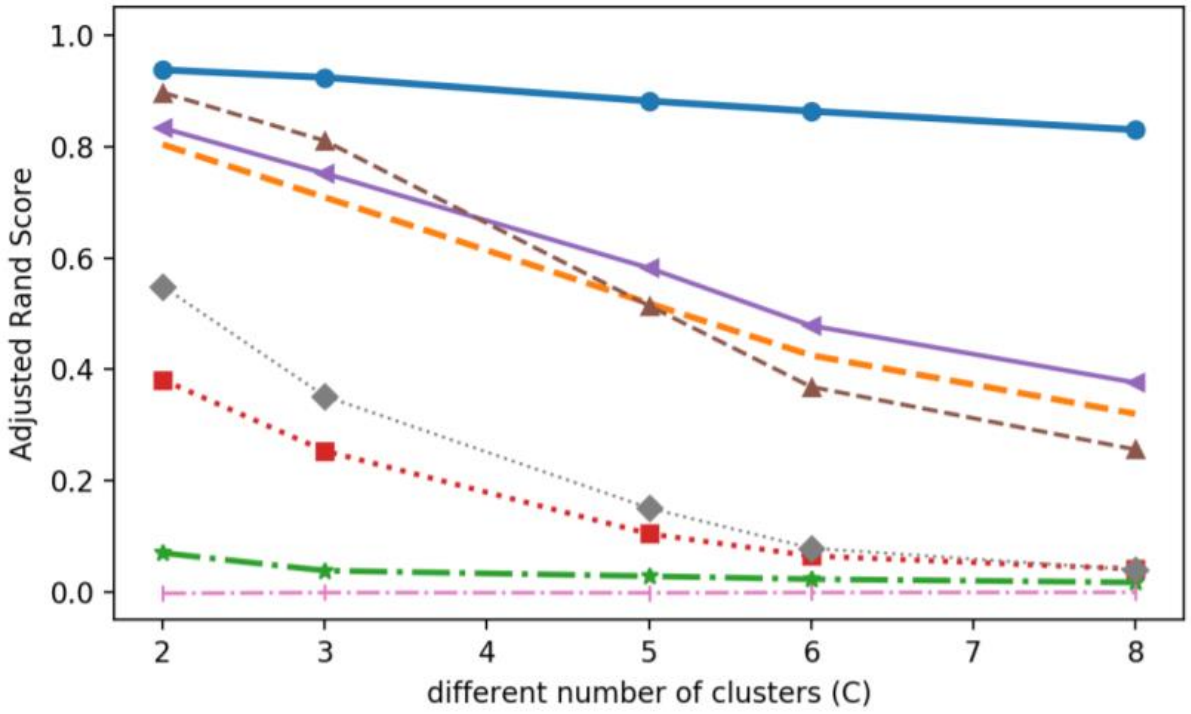
R: the number of replications; N: number of batches; L: number of sequences in each batch;
M: Length of sequence; M_p: Length of pattern, a: length of alphabet set, C: number of clusters

R = 10
N = 1080
L = {40, 120, 200}
M = {20, 40, 90}
M_p = {(2,3),(4,6),(6,9)}
a = {4, 6, 10, 20}
C = 2

ACTAATGAATCTTACCCACCATGGTCA
ATCGATACTGATCTGAATGGGGACCAT
ATCGTAGCTTAGCTATCGATTTCAATGT
TTTAGCTAATTCGATTTCGGTAGTAGTG

EXP IV – Sensitivity to number of clusters





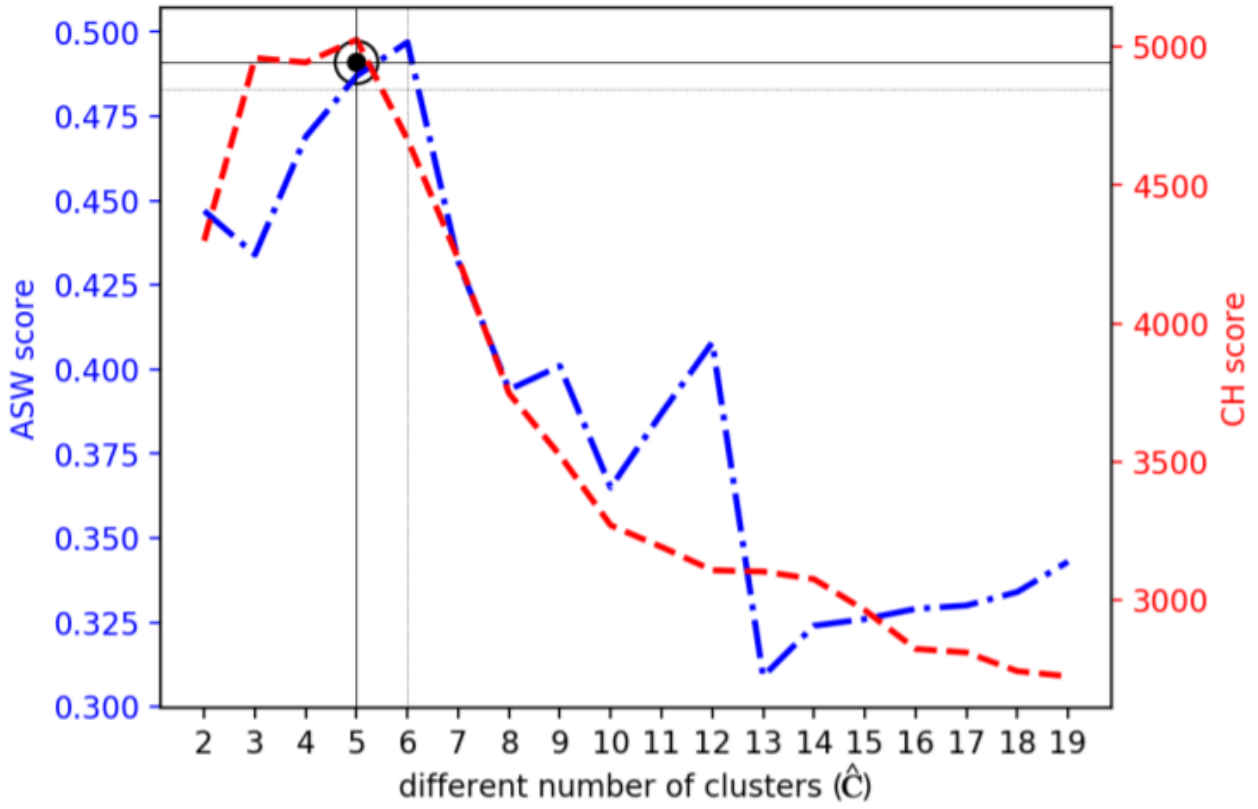
EXP V – Austrian wage data (I)

- The Austrian Wage Mobility dataset reports the wage category in successive years for the young men entering labour market.
- It consists of $L=9402$ of such workers whose gross monthly wages of successive years (ranging from $M = 2$ to $M = 32$ years) have been examined. The gross monthly wages have been divided to 6 categories, from 0 to 5, in which zero corresponds to unemployment and categories one to five correspond to the quintiles of the income distribution.
- Our prediction of number of clusters is

Pamminger, C., & Frühwirth-Schnatter, S. (2010). Model-based clustering of categorical time series. *Bayesian Analysis*, *5*(2), 345-368.

EXP V – Austrian wage data (II)

- nTreeClus predicts the possible number of clusters as 5.

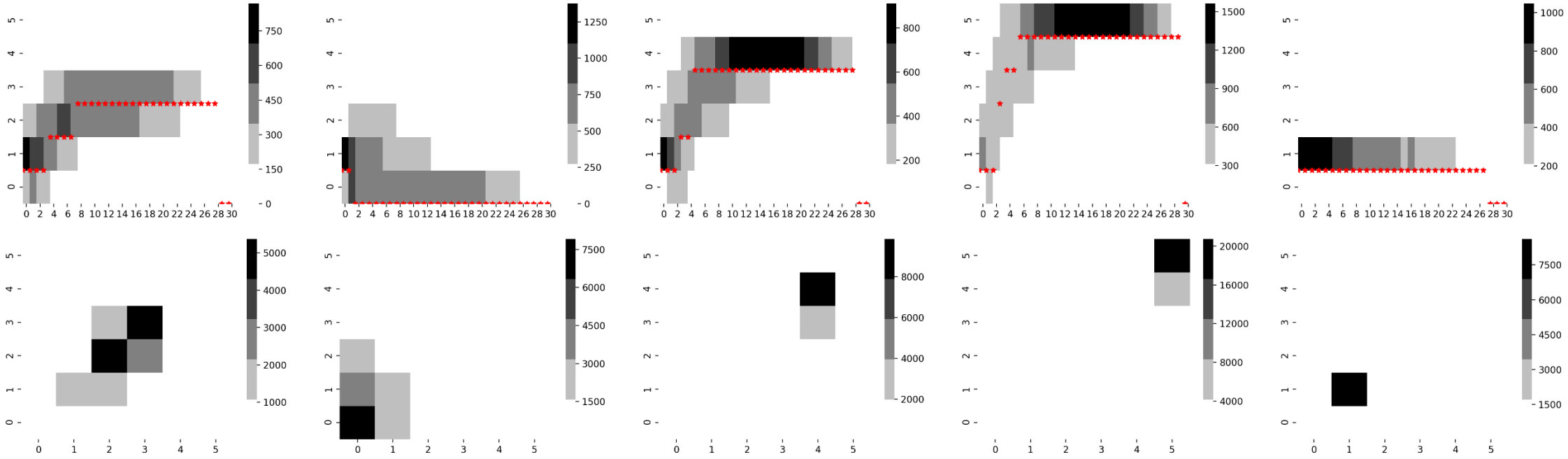


R = -
N = 1
L = 9402
M = [2, 32]
M_p = -
a = 6
C = ?

R: the number of replications; N: number of batches; L: number of sequences in each batch; M: Length of sequence; M_p: Length of pattern, a: length of alphabet set, C: number of clusters

EXP V – Austrian wage data (III)

- Here is how these 5 clusters are promoting to different job levels during their service time.
- For instance, the most left figure indicates that some of the employees always remain in level 1 and never get promoted.



EXP VI – Protein Data

Methods	External Validation				Internal Validation
	Purity	RI	ARI	F-meas	ASW
nTreeClus (DT)	1.000	1.000	1.000	1.000	0.815
nTreeClus (RF)	1.000	1.000	1.000	1.000	0.832
Levenshtein	1.000	1.000	1.000	1.000	0.950
Jaro-Winkler	0.750	0.500	-0.001	0.480	-0.003
k-mers	0.949	0.903	0.805	0.948	0.444
MHMM	1.000	1.000	1.000	1.000	-

R = -
N = 1
L = 2112
M = [80,127]
M_p = -
a = 20
C = 2

R: the number of replications; N: number of batches; L: number of sequences in each batch;
M: Length of sequence; M_p: Length of pattern, a: length of alphabet set, C: number of clusters

Concluding remarks

- Sequence mining is a need to be addressed by data scientists and some of the state-of-the-art algorithms face major hurdles including universal applicability, computational complexity, sensitivity to the position of a pattern and the length of sequences, and vulnerability to parameter setting.
- Our experimental results show that nTreeClus is robust towards parameter setting and provides computationally efficient and superior results on real and synthetic benchmark datasets with heterogeneous characteristics.
- Future work:
  - The performance of the method for classification should be explored. Furthermore, nTreeClus should be extended to empirically and theoretically investigate categorical sequences which are continuous series of elements and where the time between two of their elements is of importance - i.e. clickstream datasets.