AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Clustering Vital Sign Observations Using Unsupervised Random Forest

Dae Hyun Lee and Meliha Yetisgen

Biomedical and Health Informatics, University of Washington

Disclosure

We have no relevant relationships with commercial interests to disclose.

Introduction

Extracting patterns from time-series clinical variables is important for understanding patient progression

By considering variables that are commonly observed in EHR, we can extract patterns from most of the situation while minimizing the need of handling missing values

Considering the relationship between clinical variables could improve the quality of clustering results

Vital Signs as baseline description of patient status

In clinical dataset, variables are observed with different frequency

To extract physiologic patterns from time-series clinical observations, candidate features should be:

- Regularly measured during patient's hospital stay
- The subset of variables should have clinical meanings as a set

Vital signs are routinely checked throughout hospital units

- Vital signs are robust measurements compared to other types of clinical observation
  - Human body has mechanisms to counter-balance sudden disturbance
  - There exists dedicated interventions to stabilize vital signs when distorted significantly

Clustering

Clustering analysis is common methods to extract generalizable patterns

Clustering analysis includes

- selection of feature representation
- selection of clustering approach

Why do we introduce additional feature transformation process?

They can better represent the relationship between features

- Potential for improving the quality of clustering analysis
- Most of the distance measures were calculated based on independent contributions of feature difference

$$Minkowski\ distance\ of\ order\ p\ between\ x_i\ and\ x_j:$$

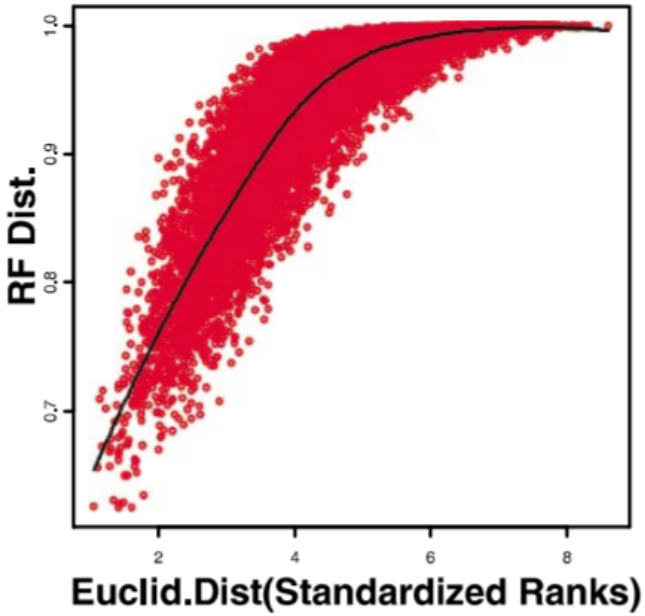
$$\left(\sum_k \left|x_{i,k} - x_{j,k}\right|^p\right)^{\frac{1}{p}}$$

Unsupervised Random Forest (URF)

Proposed by Shi et al.,

Introduce synthetic data and train a discriminator model to classify whether the instance is synthetic or original

The distance measure from URF penalizes farther instance compared to Euclidian distance



Shi T, Horvath S. Unsupervised learning with random forest predictors. J Comput Graph Stat 2006 Mar;15(1):118–38.

Why synthetic data?

Feature abstraction without additional data

- PCA, Denosing Autoencoder

Synthetic data allow us to extract characteristics from the dataset by comparing to randomness

- Generative Adversarial Network(GAN): trains the latent variables explaining the distribution of instances by introducing synthetic data
- URF : trains the conditions frequently observed in the dataset by introducing synthetic data

GAN, PCA, Denoising Autoencoder: $f(< x_1, \dots, x_d >) =< x'_1, \dots, x'_k >$

URF: $f\ (< x_1, \dots, x_d >) =< s_1, \dots, s_k > where\ s_k: \begin{cases} \bigcap_{k'} x_{k'} \geq TH_{k'} : True \\ \neg \bigcap_{k'} x_{k'} \geq TH_{k'} : False \end{cases}$

Preprocessing for URF training

Original Dataset

Instance	Feature 1	…	Feature D
Orig1	3	…	115
…	…	…	…
OrigN	3.42	…	131

+

Label
Original
…
Original

Synthetic Dataset

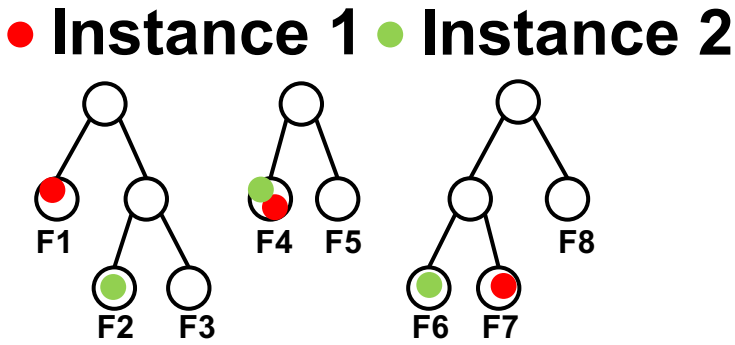
Instance	Feature 1	…	Feature D
Synth1	3	…	115
…	…	…	…
Synth N	3.42	…	131

+

Label
Synthetic
…
Synthetic

Unsupervised Random Forest

Unsupervised Random Forest and Terminal Node Assignment



	F1	F2	F3	F4	F5	F6	F7	F8
I1	1	0	0	1	0	0	1	0
I2	0	1	0	1	0	1	0	0

Similarity = 1/3, Dissimilarity = $\sqrt{1 - 1/3}$

K-medoids clustering

Instead of using centroid (within group average), k-medoids selects an instance as a representative of each cluster

1. Initialize: select k of the n data points as the medoids
2. Associate each data point to the closest medoid.
3. While the cost of the configuration decreases:
    1. For each medoid m, for each non-medoid data point o:
        1. Swap m and o, recompute the cost (sum of distances of points to their medoid)
        2. If the total cost of the configuration increased in the previous step, undo the swap
4. Go back to Step 2 if any medoids changed compared to the previous step

Kaufman L, Rousseeuw P. Clustering by means of medoids. North-Holland, 1987.

Dataset

MIMIC 3 Dataset, extracted vital signs only from patients age over 18 yrs

Vital sign

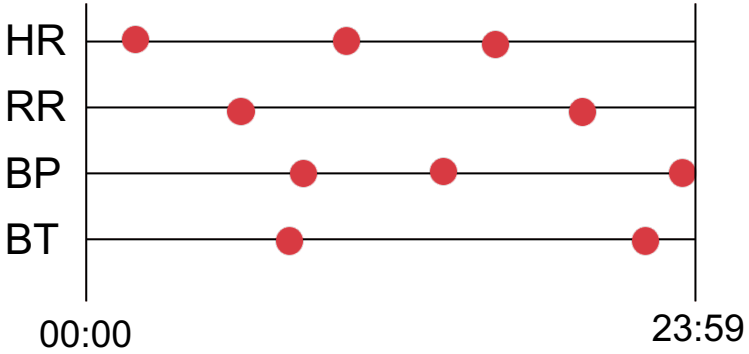
- Heart Rate
- Respiratory Rate
- Mean Arterial Pressure
- Body Temperature

The unit of analysis: vital signs from each patient-day

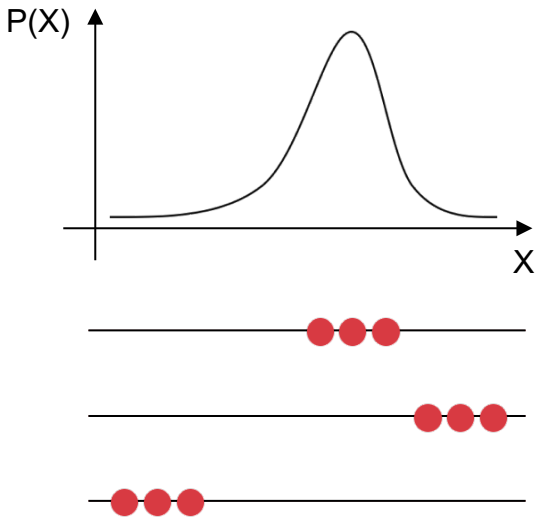
- Total 345,053 patient-day available in the dataset
- Extracted 254,716 patient-day after case imputation

Preprocessing

Patient A on Day D



Sustainment Quantifier



[Var]_GT = 1-[Var]_LT

Two-tailed P-value ([Var]_TT)	One-tailed P-value (left tail) ([Var]_LT)
High	High
Low	Low
Low	High

Summary statistics

Average, Min, Max, Standard Deviation

Patient	Day	HR_AVG	…	BT_LT
A	D	89.3		0.001

Evaluate the clustering result

Hypothesis

- High quality clustering results will show clear distinction on frequently observed patterns between expired patients and survived patients
- The evaluation should be done using a gold standard measure representing patient's severity available in EHR

Defined two quality measures for evaluation

- $Cluster\ Mortality: CM_i = \frac{\#\ of\ instances\ from\ expired\ patients}{\#\ of\ instances\ in\ the\ cluster}$
- $Intercluster\ Mortality\ Difference = \max(|CM_i - CM_j|, for\ \forall i \neq j)$

Comparative Analysis

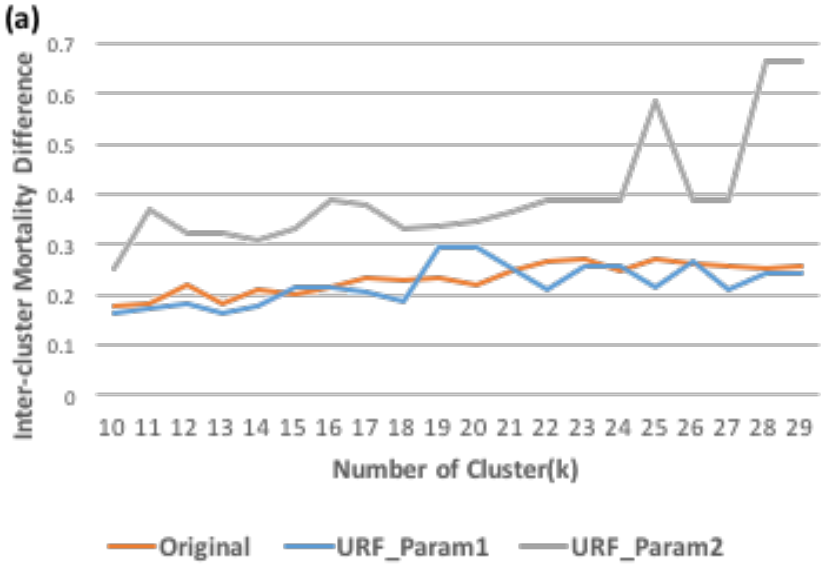
The quality of clustering result with URF-generated features vs. raw features on k-means clustering

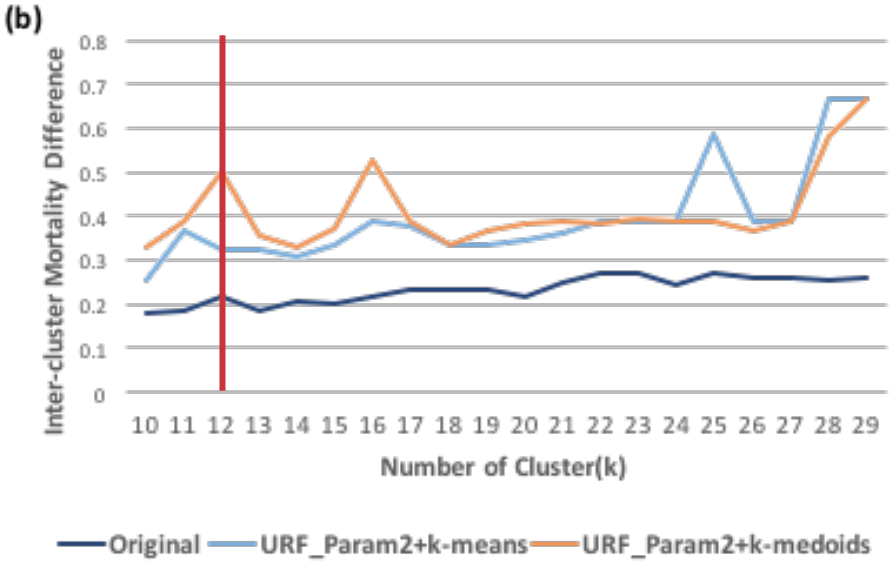
- Different URF-generated features by different hyper parameters

The quality of clustering results with different feature representation and clustering methods

- K-means & raw features
- K-means & URF features
- K-medoids & URF features

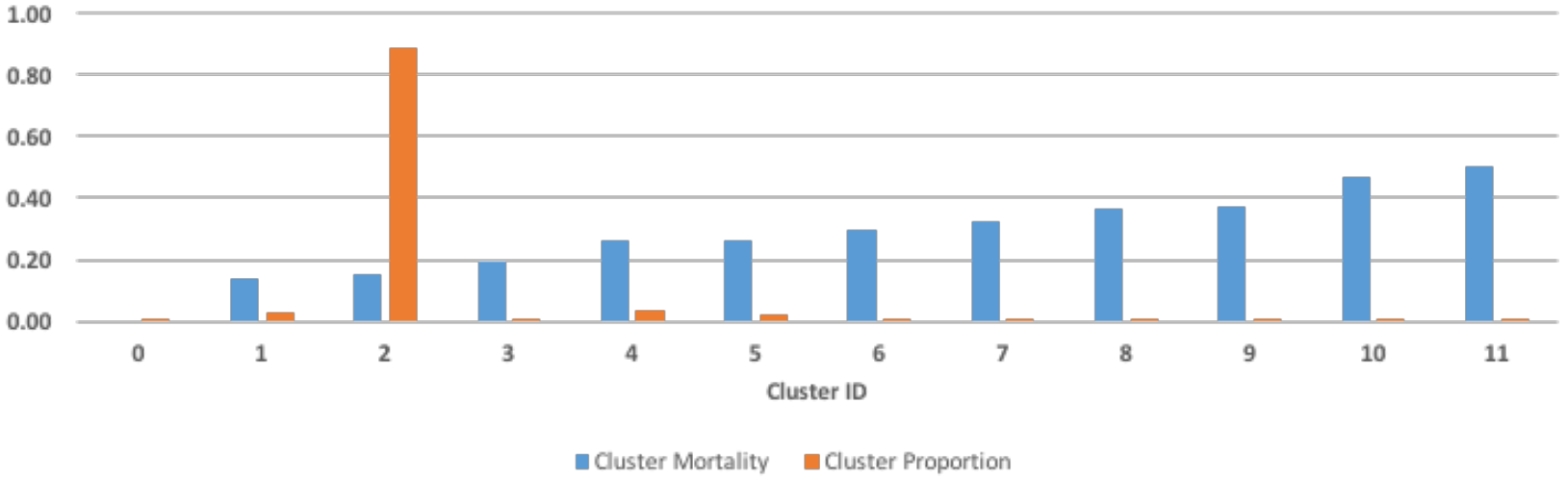
Comparative Analysis Result





Feature Representation	Number of features
Original	28
URF_Param1	20480
URF_Param2	20

Selected clustering result for further validation



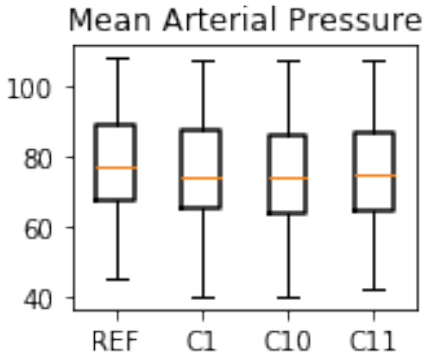
Cluster 0

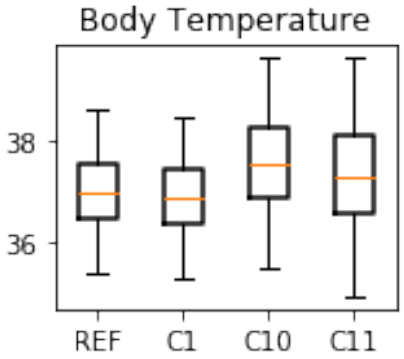
# of instances	$HR_{min} \leq 112 \wedge BP_{std} \leq 3.15$	$HR_{max} \leq 153 \wedge HR_{min} \leq 112$	$HR_{GT} \geq 0.001 \wedge HR_{min} \leq 129$	$HR_{GT} < 0.001 \wedge HR_{min} \leq 112$	$HR_{avg} > 126 \wedge HR_{max} \leq 153$	$HR_{min} \leq 101$
3	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
1	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE

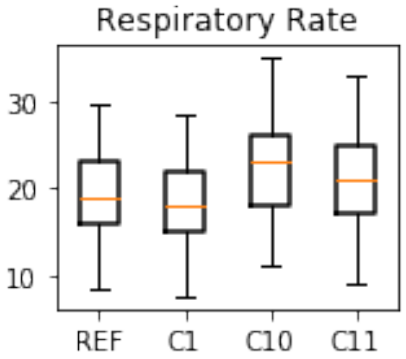
$$medoid\colon HR_{min} \leq 101 \wedge HR_{max} \leq 153 \wedge BP_{std} \leq 3.15 \wedge HR_{avg} > 126 \wedge HR_{GT} \geq 0.001$$

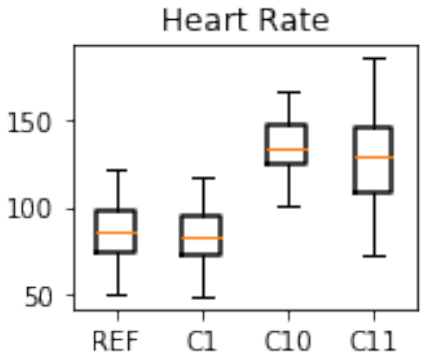
Developed tachycardia with under specific MAP variance threshold

Original feature distribution on each cluster









Medoids assertion on clusters with low and high cluster mortality

Terminal Node Condition	Medoid 1	Medoid 11
$HR_{min} \leq 112 \wedge BP_{std} \leq 3.15$	TRUE	FALSE
$HR_{min} \leq 112 \wedge BP_{std} > 3.15$	FALSE	TRUE
$HR_{max} \leq 153 \wedge HR_{min} \leq 112$	TRUE	FALSE
$HR_{max} > 153 \wedge HR_{std} > 17.8$	FALSE	TRUE
$HR_{GT} > 0.001 \wedge HR_{min} \leq 129$	TRUE	FALSE
$HR_{GT} \leq 0.001 \wedge HR_{min} \leq 112$	FALSE	TRUE
$HR_{avg} \leq 126 \wedge HR_{GT} > 0.12$	TRUE	FALSE
$HR_{avg} > 126 \wedge HR_{max} > 153$	FALSE	TRUE
$HR_{min} \leq 101$	TRUE	TRUE

Medoid 1:

$HR_{min} \leq 101$
$HR_{max} \leq 153$
$BP_{std} \leq 3.15$
$HR_{GT} > 0.12$
$HR_{avg} \leq 126$

Medoid 11:

$HR_{min} \leq 101$
$HR_{max} > 153$
$HR_{std} > 17.8$
$BP_{std} \geq 3.15$
$HR_{GT} \leq 0.001$
$HR_{avg} > 126$

Computational efficiency on URF features

From dataset, we have ~250K observations to be clustered

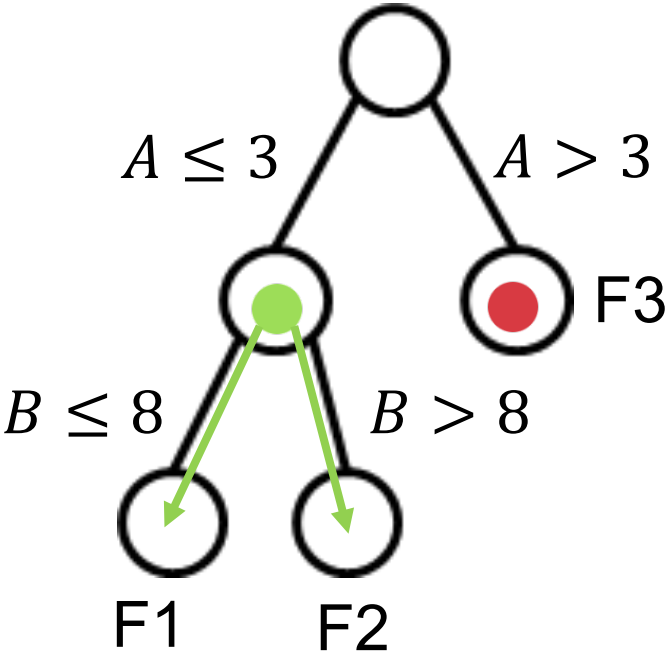
- URF transformation changes them into 56 distinct representation

Since URF representation only includes binary variables, calculating similarity between URF representation is faster than calculating Euclidian distance between raw representation(float)

→ Instead of calculating $\binom{250K}{2}$ Euclidian distances, calculate $\binom{56}{2}$ hemming distances

Further Work

Representing instances with missing data using URF



Instance	A	B
I1	5	7
I2	2	?

Instance	F1	F2	F3
I1	0	0	1
I2	0.5	0.5	0

Conclusion

Using URF as preprocessing approach were able to grab richer representation from the original feature

URF representation along with k-medoids clustering showed computational efficiency compared to other approaches while outperformed on the evaluation criteria

URF representation allowed us interpret clusters in terms of original features

Thank you!

Email me at:
dhlee4@uw.edu