A k-means approach to clustering disease progressions

Duc Thanh Anh Luong          Varun Chandola
Department of Computer Science & Engineering
University at Buffalo

IEEE ICHI 2017
August 26, 2017
Outline

• Motivation
• K-means approach
• An application for Chronic Kidney Disease
• Generating patient-specific disease profiles
Motivation

• Find subgroup of patients have similar disease progression
• Identify the underlying mechanism of the subgroup
• Provide better treatment for each subgroup
Motivation

• Different patients have different disease progressions
• Consider the case of Chronic Kidney Disease

Are there few general trends of disease progressions?
Can we group patients by their progressions into few groups?
Motivation

Trajectories of 500 patients

Cluster 1
422 patients (5.91%)

Cluster 2
1061 patients (14.86%)

Cluster 3
980 patients (13.72%)

Cluster 4
1042 patients (14.59%)

Cluster 5
1070 patients (14.98%)

Cluster 6
686 patients (9.61%)

Cluster 7
825 patients (11.55%)

Cluster 8
678 patients (9.49%)

Cluster 9
326 patients (4.56%)

Cluster 10
52 patients (0.73%)

Trajectories after being clustered
Clustering problem and k-means algorithm

- Cluster a set of data points into k clusters
- Can be solved by K-means approach

K-means approach

Data object → Patient disease progression → Distance metric → Centroid → Regression line

- eGFR (estimated glomerular filtration rate) over time:
  - Top graph: eGFR values with a peak around day 1000.
  - Bottom right graph: eGFR values with a peak around day 900.
  - Bottom left graph: eGFR values with a peak around day 400.

Centroid:
- Red line on the graphs representing the centroid.
**K-means approach**

Initial step:
- randomly assign patient into k clusters

Update step:
- Perform **regression** for each cluster to obtain “centroid”

Assignment step:
- Assign patient to the cluster that has **closest** centroid

**Diagram:***
- **No patient move to another group?**
  - Yes: **End**
  - No: Repeat Update and Assignment steps
**Dataset & Preprocessing**

DARTNet patients (n = 69,817)

- Invalid birth year and sex value (n = 6,418) → Excluded
- Invalid data records (n = 9) → Excluded

"Preprocessed" DARTNet patients (n = 63,209)

- Observation duration < 1 year (n = 5,285) → Excluded

Final CKD cohort (n = 7,142)

- Having eGFR values less than 60 for more than three months (n = 29,585) → Excluded
- Number of serum creatinine records < 1 (n = 181) → Excluded
- Number of serum creatinine records < 10 (n = 17,158) → Excluded
Clustering result

Cluster 1
422 patients (5.91%)

Cluster 2
1061 patients (14.86%)

Cluster 3
980 patients (13.72%)

Cluster 4
1042 patients (14.59%)

Cluster 5
1070 patients (14.98%)

Cluster 6
686 patients (9.61%)

Cluster 7
825 patients (11.55%)

Cluster 8
678 patients (9.49%)

Cluster 9
326 patients (4.56%)

Cluster 10
52 patients (0.73%)
Demographic distribution in clusters

- **Cluster 1**: 14.59% (1042)
- **Cluster 2**: 14.86% (1061)
- **Cluster 3**: 13.72% (980)
- **Cluster 4**: 14.98% (1070)
- **Cluster 5**: 9.61% (686)
- **Cluster 6**: 11.55% (825)
- **Cluster 7**: 9.49% (678)
- **Cluster 8**: 4.56% (326)
- **Cluster 9**: 0.73% (52)
- **Cluster 10**: 5.91% (422)

- **Gender Distribution**
  - Male
  - Female

- **Age Distribution**
  - Cluster 1: 4.56%
  - Cluster 2: 5.00%
  - Cluster 3: 7.50%
  - Cluster 4: 10.00%
  - Cluster 5: 2.50%
  - Cluster 6: 5.00%
  - Cluster 7: 7.50%
  - Cluster 8: 10.00%
  - Cluster 9: 2.50%
  - Cluster 10: 5.00%
Other clinical markers

- Albumin-to-creatinine ratio
- Hemoglobin A1c
- High Density Lipoprotein
- Alanine Aminotransferase
- Phosphorous
- Fasting Blood Glucose
- Low Density Lipoprotein
- Aspartate Aminotransferase
- Parathyroid Hormone
- Non-Fasting Blood Glucose
- Triglyceride
Generating patient-specific disease profiles

Generating patient-specific disease profiles

- **cluster's trajectory**
- **individual predicted trajectory**
- **upper and lower limit**
- **actual eGFR value**
Conclusion & Future Work

• Clustering disease progressions – k-means approach
• Generating individual prediction – Gaussian processes

• Extend the approach to cope with multiple clinical markers
• Give quantitative evaluation of clusters
  • Tightness
  • Separation
Thank you