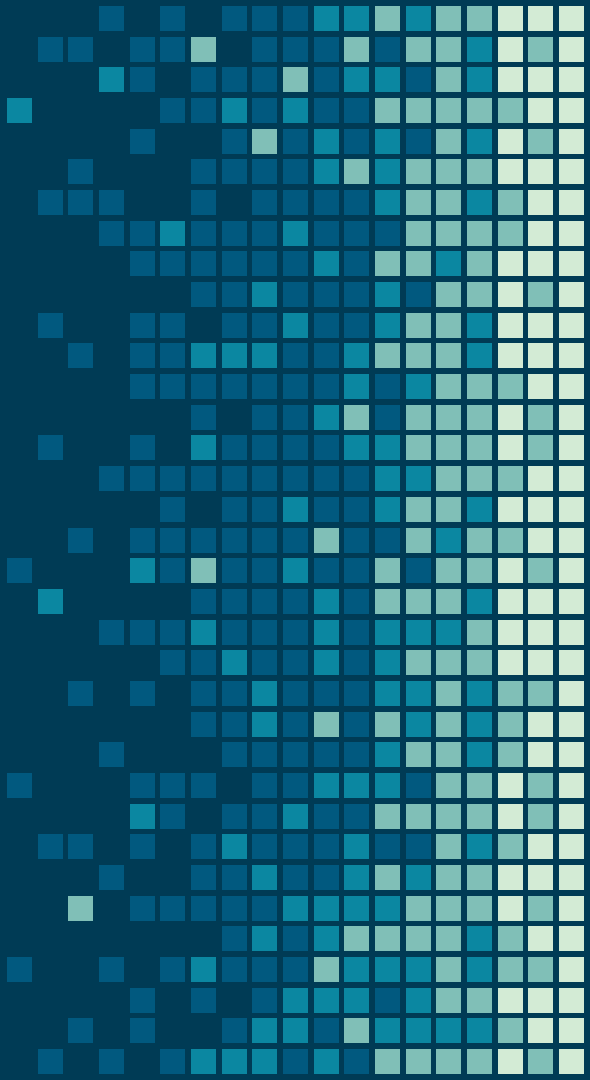
A K-Means Approach to Clustering Disease Progressions

Presented by: Gentry Atkinson



Conference Paper:

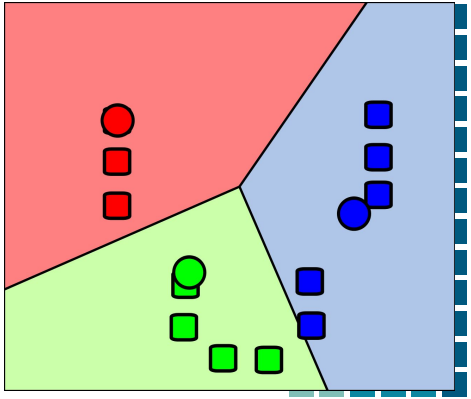
- "A k-means approach to clustering disease progressions"
- Duc Thanh Anh Luong, Varun Chandola. Both with the University of Buffalo
- 2017 IEEE International Conference on Healthcare Informatics

Summary:

- Identifying patient subpopulations using electronic records can improve care.
- A new time-series distance measure is proposed.
- 10-means clustering is applied to patient data to identify 10 unique cluster.
- Predicted patient outcomes are comparable to older techniques which use more data.

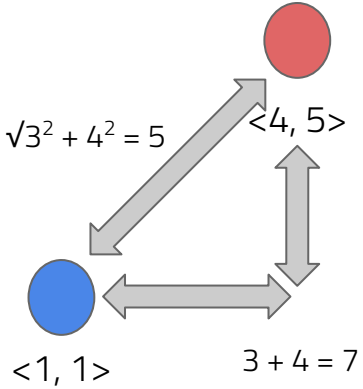
Background- K Means:

- Clustering uses unsupervised learning to identify groups in data sets based on some **distance measure**.
- K-Means is one of the oldest and most relied upon clustering algorithms. [1967]
- Guesses k centroids in the data and then refines to minimize total distances from points to centroids. Iterates to stability.



Background- Distance Measures:

- Quantifies the similarity or dissimilarity of two data points.
- Very application specific.
- Some common examples:
  - Euclidean
  - Manhattan
  - Cosine
  - Jaccard Coefficient



Background- CKD:

- Chronic Kidney Disease affects 753 million people worldwide.
- Affects a wide range of patients types by gender, age, and other factors.
- Measured with eGFR: the estimated glomerular filtration rate, which measures the flow rate of filtered fluid moving through the kidneys.

Related Work:

- Another distance measure Dynamic Time Warping handles time series data well but has difficulty with sparse and misaligned data. TW Liao 2005.
- Probabilistic Subtyping Model explain individual disease progression but is more sensitive to hyper-parameters. Luong et al. 2017

Modified distance measure:

- Assumes that time series can be approximated using a weighted sum over a collection of splines (a little smoothed out function).
- Distance from patient to cluster centroid is equal to the sum of square distances between the patients data and the centroids predicted progression, based on eGFR.

The dissimilarity between a patient i and a cluster k can be measured using the sum of square difference between his/her actual lab measures and its representative progression of cluster k, and is computed as $\left\|\mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(k)}\right\|_2^2$.

Modified K-Means:

- Minimizes error (distance) through an iterative process.
  - Guesses a centroid
  - Adjusts centroid to minimize distance
  - Repeats until stable.
- K=10 was chosen experimentally.

Initialization:
for $i \in 1, \cdots, N$ **do**
    randomly assign patient i to a cluster
end for
repeat
    Update step:
    for $k \in 1, \cdots, K$ **do**
        compute $\boldsymbol{\beta}^k$ which minimizes
        $\sum_{i=1}^{N} \sum_{k=1}^{K} z_{i,k} \left\| \mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(k)} \right\|_2^2$
        using the method of least squares
    end for
    Assignment step:
    for $i \in 1, \cdots, N$ **do**
        set $z_{i,k} = 1$
        where $k = \arg\min_j \left\| \mathbf{x}_i - \Phi(\mathbf{t}_i)\boldsymbol{\beta}^{(j)} \right\|_2^2$
        and set $z_{i,j} = 0$ for $j \neq k$
    end for
until Convergence ($\{\mathbf{z}_i\}_{i=1}^{N}$ does not change)

The Data:

- Public data set collected by DARTNet.
- 69,817 anonymous patients with various levels of kidney damage.
- eGFR is normalized to account for patient age, sex, and race.
- Includes many patients which do not meet the clinical criteria for CKD.

Data Preprocessing:

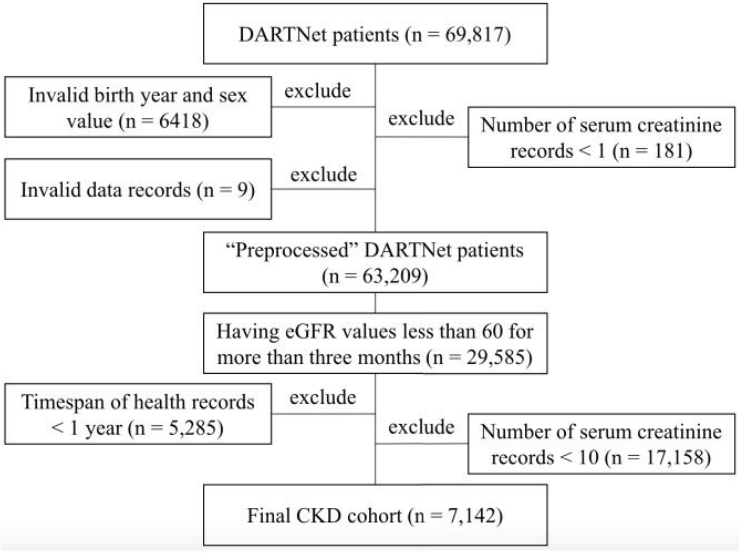


Fig. 1: Preprocessing procedure to obtain CKD cohort

- The study only retained patients who:
  - Met criteria for CKD
  - Had one year of eGFR values.
  - Had >10 serum creatinine measures.
- Final set was 7,142 records.

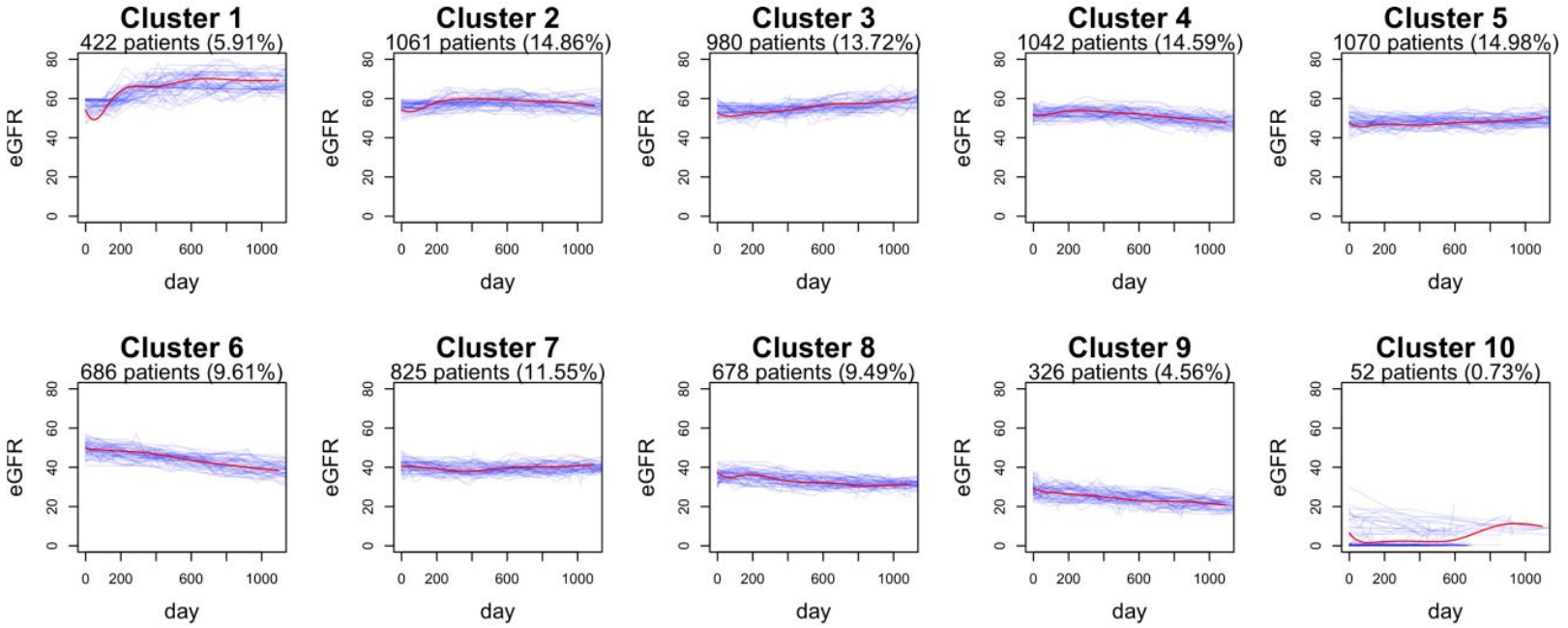
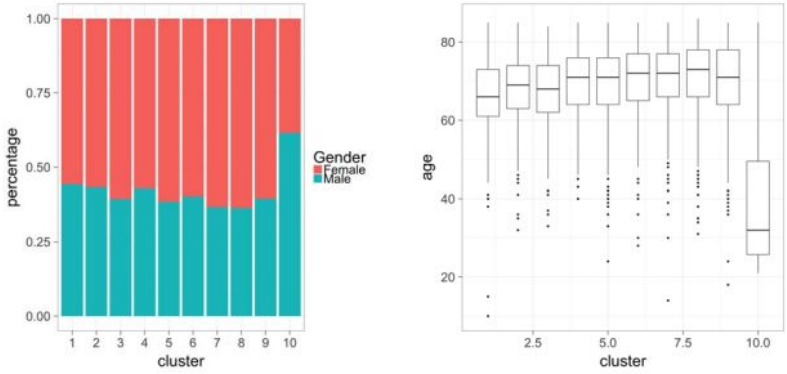


Fig. 2: Clustering output of k-means algorithm applied to the CKD dataset. In blue are the raw patient observations and in red are the representative disease profiles. Number of patients and its proportion in CKD cohort are given for each cluster.

Results:

- Clusters 1-3 correspond to patients with slightly improving disease progressions. **34.9% of population.**
- Clusters 4-9 mostly contain patients in stage 3 and 4 of CKD. **64.8% of population.**
- Cluster 10 contained patients with severe kidney damage. **0.73% of population.**



(a) The distribution of gender for each cluster

(b) The distribution of baseline age for each cluster

Fig. 3: Demographic distribution for each cluster

Effectiveness of Predictions:

- Gaussian Process Regression was used to prepare patient-specific prediction of eGFR at a future time using the representative profile of their cluster.
- This method is compared to a model tailored for time-series, eGFR data called *Monitor.* Diggle, Sousa, and Asar 2015.
- K-Means achieves similar accuracy to *Monitor* with a generalized model that depends on less information, as *Monitor* also uses information such as baseline age and gender.

Model	Overall RMSE
Modified k-means	8.98
Monitor	6.66

TABLE I: Comparison of prediction error in SRFT dataset

(RMSE is Root Mean Square Error)

Discussion:

- The authors observe that patients with similar disease progressions are likely to share common disease mechanics. Identifying clusters might reveal specific disease subtypes.
- K-Means runs in **near** linear time which is very efficient compared to other methods in this field.
- Generalized models that can automatically review Electronic Health Records are important for understanding disease progression in patients.
- Distance measures are always an real challenge to clustering. This contribution is meaningful on its own.

Conclusion:

- K-Means is shown to be a viable method for grouping disease progressions.
- The clusters don't have to be meaningful on their own for them to be useful to researchers.
- This distance measure and clustering method can be applied easily to other diseases.

Questions or Comments?