Clustering Patients with Tensor Decomposition

Matteo Ruffini [1] Ricard Gavaldà [1] Esther Limón [2]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain

[2]Institut Català de la Salut, Barcelona, Spain

Overview

Task: to provide an automated and efficient method to segment patients in groups with similar clinical profiles.

1. Similar patients $\rightarrow$ Similar cares.
2. Find recurrent comorbidities.
3. Assigning and planning resources: drugs and doctors.

Overview

Task: to provide an automated and efficient method to segment patients in groups with similar clinical profiles.

1. Similar patients $\rightarrow$ Similar cares.
2. Find recurrent comorbidities.
3. Assigning and planning resources: drugs and doctors.

Dataset: all hospital admissions in Catalonia in 2016 ($>$ 1 Mln records). Each row is a visit: up to 10 diagnostics in ICD-9 format.

ICD-9 EHR

In ICD code, to each disease is associated a number
Records: list of patients with their diseases $\rightarrow$ patient-disease matrix.

	Diseases
Patient 1	820, 401
Patient 2	401, 278,
Patient 3	560, 820, 278

	820	**401**	**278**	**560**
Patient 1	1	1	0	0
Patient 2	0	1	1	0
Patient 3	1	0	1	1

ICD-9 EHR

In ICD code, to each disease is associated a number
Records: list of patients with their diseases → patient-disease matrix.

	Diseases
Patient 1	820, 401
Patient 2	401, 278,
Patient 3	560, 820, 278

	820	**401**	**278**	**560**
Patient 1	1	1	0	0
Patient 2	0	1	1	0
Patient 3	1	0	1	1

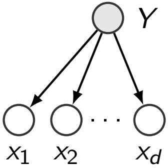
Objective: cluster the rows of the patient-disease matrix.
Sparse and high dimensional data.

Standard methods: *k-means, k-medioids, single linkage...*
Distance-based: poor performances on high dimensional sparse data.

Modeling strategy

Data is modeled as a mixture of independent Bernoulli variables

- Latent state $\rightarrow$ Medical status of a patient.
- Observed diseases depend on the patient status.
- Once in a status, diagnostics are independent.



Main advantages

- No distance required.
- Generative model $\rightarrow$ clear interpretation.
- Clustering is performed via MAP assignment.

Learning procedure: method of moments

1. Retrieve from data estimates of the *moments:*

$$
\begin{aligned}
M_1 &= \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \ \in \mathbb{R}^d \\
M_2 &= \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \otimes \mu_i \ \in \mathbb{R}^{d \times d} \\
M_3 &= \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \otimes \mu_i \otimes \mu_i \ \in \mathbb{R}^{d \times d \times d}
\end{aligned}
$$

   Where $M = [\mu_1, ..., \mu_k]$ and $\omega = (\omega_1, ..., \omega_k)$ are the unknown centers of the mixture and the mixing weights.

2. Obtain mixture's parameters with tensor decomposition on the moments:

$$\mathcal{TD}(M_1, M_2, M_3) \rightarrow (M, \omega)$$

Main challenge:
To estimate the moments from data; we used an approximated approach.

Experimental results - two subset datasets

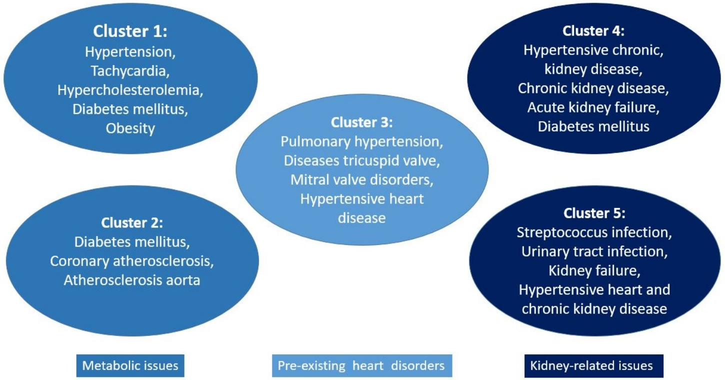
We focus on two subsets of our dataset:

1. **Heart Failure Dataset**:
   Patients having diagnostic 428 in the ICD-9 code (Heart Failure).

2. **"Tertiary" Dataset:**
   Patients with serious diseases to be treated in top hospitals.

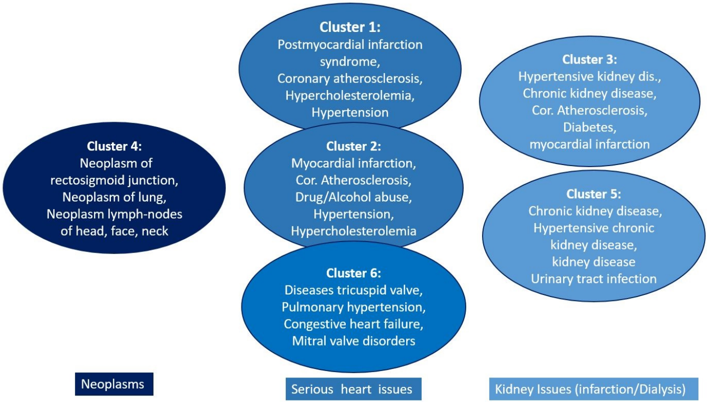
Both contain around 20000 patient records.

Heart Failure Dataset - Content of the clusters



Cluster ID:	1	2	3	4	5
Size:	7290	2915	4408	2936	5533

“Tertiary” Dataset - Content of the clusters



Cluster ID:	1	2	3	4	5	6
Size:	4892	3982	1043	3133	819	2442

Clustering Patients with Tensor Decomposition

Matteo Ruffini [1] Ricard Gavaldà [1] Esther Limón [2]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain

[2]Institut Català de la Salut, Barcelona, Spain