FALCON : An active clustering solution to anticipate aging trajectories

Blase V, Aligon J, Garouani M, Ader I & Teste O

Institut de recherche en Informatique de Toulouse
Institut RESTORE

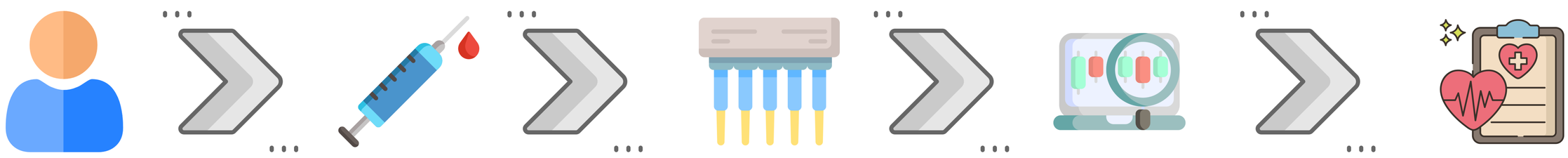
Scientific context

Unhealthy aging has become a **public health concern** due to the increase of life expectancy

World Health Organization
✓ 5 functions as indicators of unhealthy aging
- Cognition
- Sensory
- Vitality
- Locomotion
- Psychosocial

FRAILTY SYNDROME :
Decline of functions leading to an increased **vulnerability** of individuals

Identify biomarkers of frailty syndrome



Blood sample → Identification of **biomarkers** → Aging **trajectories prediction** → Care adapted to the **correction** of the aging trajectory

Hypothesis : A combination of biomarkers can predict aging trajectories

First approach

Statistical analysis

A. Volunteers selection (n=731) **B.**

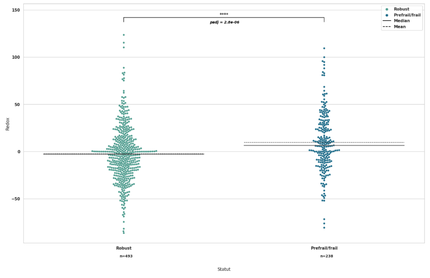


Table 2. Associations between Redox and primary clinical information			
	β	S.E.	p
Model 1			
Age	0,1594	0,0630	0,011
Sex (women)	-6,3722	2,3230	0,006
IMC	0,6427	0,2790	0,095
Model 2			
Age	0,1373	0,0690	0,047
Sex (women)	-7,3693	2,4760	0,003
BMI	0,6382	0,2960	0,057
CCI	0,0839	3,1830	0,979
Physical activity	-0,0046	0,2960	0,004

Figure 1. Statistical analysis between biomarkers and frailty syndrome (A) 731 volunteers from INSPIRE-T human cohort were included and classified according to their frailty status (robust, pre-frail, frail). Shows significant difference between robust and frail according to Redox biomarker **(B) Associations between Redox biomarker and primary clinical information** related to aging (Age, Sex, BMI, Physical activity)

Machine learning methods

Not working well for many reasons :

- Not **enough** volunteers
- Too many **information** in data variables
- Need more **complex** models
- **Frailty syndrome** is not suitable for the youngest (<65 years old)

Active learning clustering

Little Experiment

- 86 old
- BMI = 20
- Low physical activity
- Hypertension
- Little strength
- Bad vision

- 29 old
- BMI = 20
- High physical activity
- No disease
- Lots of strength
- Bad vision

- An expert can say that **both individual** are in **different clusters**
- A model can retain BMI and bad vision in order **to cluste**r these two individuals **together**.

Methodology

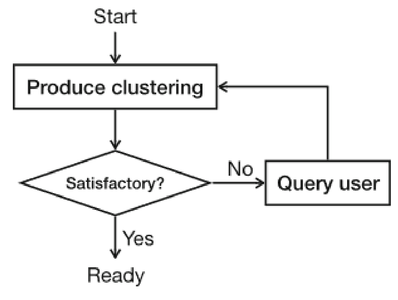
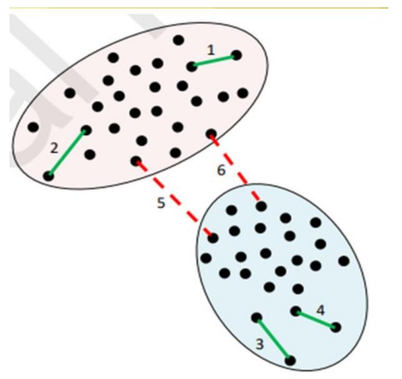


Figure 2. Illustration of the processus of Active Clustering An **initial clustering** is produce by the model. The user is ask if he is **satisfy** with it. If yes, the process end and the result is given. If no, **query is given** to the user which answer by **constraining the model** with **user-provided knowledge** in the form of **pairwise constraints**. See Fig.3

Figure 3. Illustration of the processus of pairwise constraints
When a query is given to the user, the user must answer with **pairwise constraints**, typically expressed as **Must-Link (ML) and Cannot-Link (CL)** relations, to enforce that certain pairs of data points **must belong** to the same cluster or **must be separated** into different clusters, respectively



FALCON : An Effective and Robust method for Active Clustering with Pairwise Constraints

Algorithm description

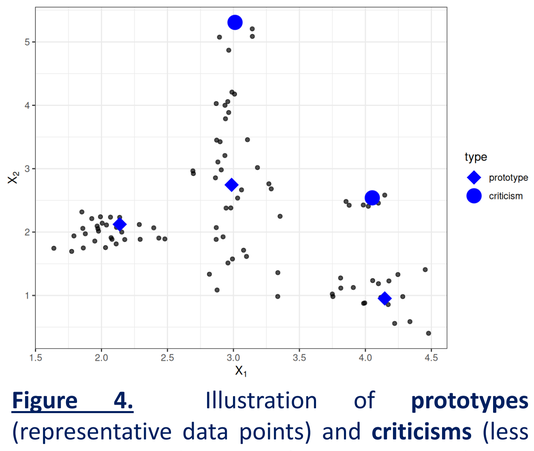


Figure 4. Illustration of **prototypes** (representative data points) and **criticisms** (less representative data points) on a toy example.

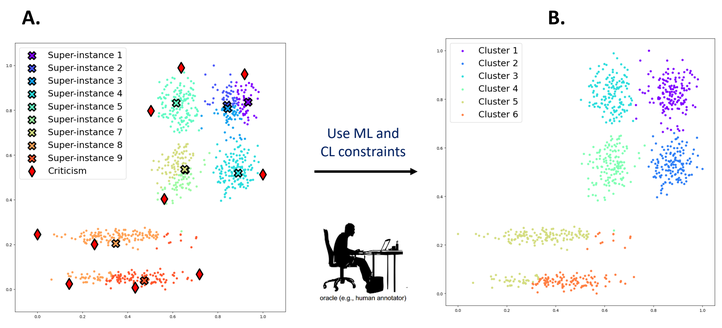


Figure 5. (A) **Super instances** (prototypes neighborhood) represents a group of data points that are assumed to belong to the same cluster. (B) Super instances are **merged** in order to form **cluster**

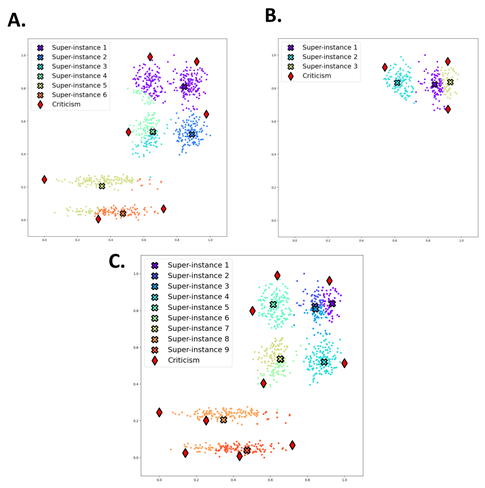


Figure 6. (A) Super instances with the **most criticism** in it is chosen for **refinement**. (B) **Refinement** is done by computing **new prototypes** and therefore super-instance in the **local super instance**. (C) New super instance are **added to the pool of global super instance** for a merging session

Experimental setting

- ❑ We compare FALCON with **two state of the art competitors** : COBRAS and ACDM and a lower baseline : Kmeans++
- ❑ We compute experiment in **28 datasets** from COBRAS and ACDM papers (UCI dataset).

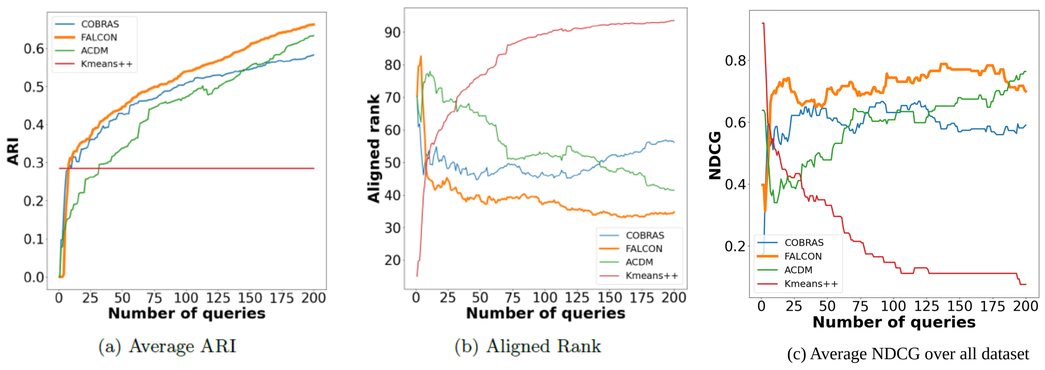
Results

RQ1 : How effective is Falcon in active clustering pairwise constraint?
RQ2 : How does FALCON perform with a small budget?
RQ3 : How does FALCON operate across different datasets size?

The Average Rand Index (ARI) computes the number of pairwise agreements and disagreements between the actual labels and those predicted by the model (RQ1)

Aligned Rank (AR) encapsulates the number of FALCON victories compared with other models. A low score is preferable

NDCG-3 evaluates a given ranking order against an ideal ranking, focusing on the top 3 positions



(a) Average ARI (b) Aligned Rank (c) Average NDCG over all dataset

FALCON consistently outperforms existing methods, not only in terms of **clustering quality** but also in its ability **to maximize the benefit of each user interaction**. This is a significant advantage in **real-world applications** where reducing user interaction is crucial.