Development of machine learning algorithms for health data analysis

DJIBEROU MAHAMADOU Abdoul Jalil[(1)], *ANTOINE Violaine*[(1)], *CHRISTIE Gregory*[(2)], *MORENO Sylvain*[(2)]
[(1)] *LIMOS*, [(2)] *Simon Fraser University – Canada*

Objectives

1. Determining new life factors that impact on dementia (Ex. Alzheimer).
2. Development of machine learning algorithms to handle categorical data.
3. development of machine learning algorithms to handle mix data (numeric and categorical).

Introduction

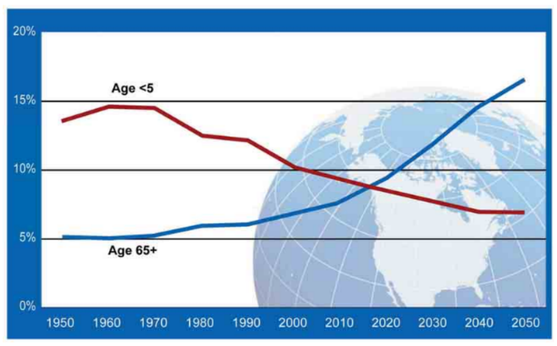

Figure 1: Projection of young and older adults [1]

- 131 million of persons are expected to be living with dementia by 2050 [2].
- Economic impact: 1 trillion total worldwide cost of dementia [2].
- A need of understanding dementia.

Background

Clustering is a family method of machine that aims to group a set of objects based on their similarity and dissimilarity. Given a number of clusters in:

- Crisp clustering, a data point must belong to at most one.
  - ▷ Ex. K-Means
- Fuzzy clustering, a data point could belong to many clusters.
  - ▷ Ex. Fuzzy C-Means
- Probabilistic clustering, certain models (Gaussian, Poisson) are used to cluster data.
  - ▷ Ex. Expectation-Maximization
- Evidential clustering, generalize crisp, fuzzy and probabilistic clustering.
  - ▷ Ex. Evidential C-Means [3]

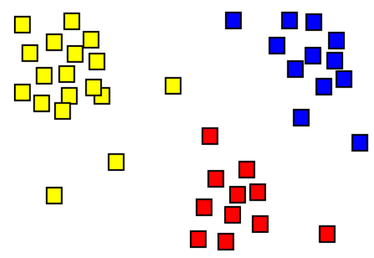

Figure 2: Example of 3 clusters

Methodology

- State-of-the-art.
- Defining dissimilarity measures.
- Defining clusters prototypes.
- Optimize the objective function.
- Experimentation.

Mathematical Section

- Objective function
  The evidential c-means (ECM) algorithm searches for a partition called credal M and the set of prototypes V that minimize intra-cluster variance:

$$\begin{cases} J_{ECM}(M,V) = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}, \\ s.t \sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1, \quad \forall i = \{1, \dots n\} \end{cases} \tag{1}$$

  With α: the weighting exponent for cardinality, β: the weighting exponent for the fuzziness, δ the distance to the empty set.
- Optimization process
  - ▷ NP-Hard problem.
  - ▷ Fix M, and optimize V.
  - ▷ Fix V, and optimize M.

Example of evidential clustering application

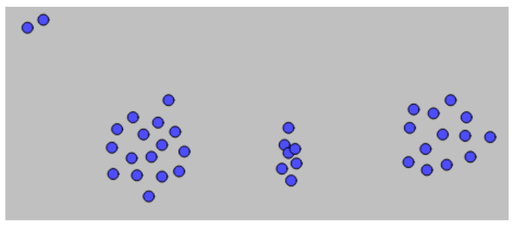

Figure 3: Data points to cluster

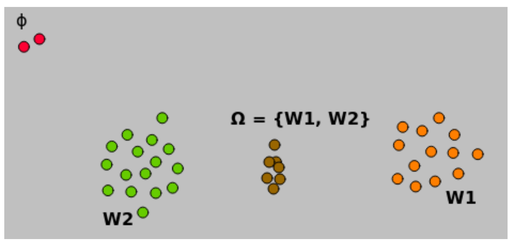

Figure 4: Results of ECM

Real-world applications:

- Green class: persons with Alzheimer.
- Orange class: persons without Alzheimer (normal).
- Brown class: persons that begin to have Alzheimer's symptoms.
- Red class: persons with lost of memory but not due to Alzheimer.

Challenges with categorical data

Let's consider the follow data set:

Sex	AgeGroup	MemoryLost	RepeatedQuestionsOrStatements
M	35-45	No	No
F	65-75	Yes	Yes
M	55-65	Yes	No
F	45-55	No	No

- How to define clusters prototypes ? (ex. Barycenter with numeric data)
- How to define distance between data points and clusters ? (ex. Eucludian distance with numeric data)

Contributions

Our paper 'Evidential clustering for categorical data' is accepted for publication at the international conference FUZZ-IEEE. The main contributions are the following:

- New cluster prototypes representation.
- New dissimilarity measure between clusters and data points.

Conclusion & Perspectives

- Dementia remains a big challenge in the world.
- Machine learning approach to determine life factors that affect dementia.pOSTER
- Development of new robust methods that can handle numeric and categorical data.
- Perspectives
  - ▷ Defining new dissimilarity measure (the proposed one gives crisp cluster centers).
  - ▷ Mix evidential clustering to handle categorical and numeric data.

References

[1] United Nations. World Population Prospects: The 2010 Revision.

[2] Alzheimers Disease International.
September 2016.

[3] M. Masson and T. Denœux.
ECM: An evidential version of the fuzzy c-means algorithm.
Pattern Recognition, 41(4):1384–1397, 2008.

Acknowledgments

- This thesis is supported by the Agence Nationale de la Recherche of the French government throught the program "Investissements d'Avenir"(16-IDEX-0001 CAP 20-25).

Contact Information

- Web: http://www.limos.fr
- Email: abdoul_jalil.djiberou_mahamadou@uca.fr
- Phone: +33 (0) 473 40 53 70