Pretopology-based Clustering for Mixed Data

Soufian Ben Amor[1], Maxence Choufa[2], Clement Cornet[2], Sonia Djebali[2],
Guillaume Guerard[2], Loup-Noé Lévy[1,3], Hai Tran[3]

[1] LI-PARAD Laboratory EA 7432, Versailles University, 55 Avenue de Paris, 78035, Versailles, France
{FirstName.LastName}@uvsq.fr

[2] Léonard de Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France
{FirstName.LastName}@devinci.fr

[3] Energisme, 88 Avenue du Général Leclerc, 92100, Boulogne-Billancourt, France
{FirstName.LastName}@energisme.com

Mots-clés : *Pretopology, Clustering, Mixed Data, Machine Learning*

1 Introduction

The energy performance of buildings represents a major issue of the 21st century. Many solutions have been discussed to improve buildings' energy performance [1, 4], but the actions to take differ from one building to another. In other words, current solutions are built on a case-by-case basis and cannot be extrapolated easily. Indeed, it is difficult to find generic solutions due to their complexity and heterogeneity.

By placing buildings in groups and subgroups, one can define relevant energy optimization recommendations without auditing each building individually. Because initial labels are not always defined, clustering is relevant in our case. Since we seek for intrinsic similarities between groups and subgroups, hierarchical clustering is needed. Buildings are described with mixed data. They include numerical data such as surface or number of floors, and categorical data like types of heating or insulation materials. Few clustering algorithms exist for mixed data, and even fewer are hierarchical. In this article, we present a method for the hierarchical clustering of mixed data based on pretopology.

2 Our Approach

Few were developed for mixed data compared to the number of clustering algorithms made for either numeric or categorical data. Roughly, research works in this domain modified the existing clustering algorithms to make them perform well on mixed data. From a survey by Ahmad et al. [2], we can distinguish four main types of clustering on which mixed data works are based : *partitional*, *hierarchical*, *model-based* and *neural-network based* clustering. The main differences between the hierarchical clustering algorithms for mixed data are the similarity measure used to compute the similarity matrix and the method of clustering applied on it. But the relevance of this matrix depends on the definition of the similarity between two mixed datapoints, which is hard to grasp. Other minor types of clustering exist for mixed data, but very few of them uses our approach.

Our approach is based on *pretopology*, which can be considered as a mathematical tool for modeling the concept of proximity It allows to extract, organize, and cluster data into homogeneous groups and to gain knowledge from the emerging structure of the population.

Lévy and al. [4] proposed a proof of concept pretopology-based clustering on time series. The algorithm has been completed. It's applied to various datasets with various measures. Then, it is compared to other clustering methods considering various cluster analysis metrics.

3 Preliminary Results

The first results provide good insight into the efficiency and quality of the proposed method. Using a reduction of dimensions, namely FAMD then the pretopology-based clustering, we compare with two mixed-data clustering methods (K-Prototype and Kamila) and K-means on the FAMD result [2, 3]. To evaluate the clustering, several internal indices have been used to describe the quality of our clustering [5]. The CH-index is the ratio of between-clusters and within-clusters dispersions and should be maximized. The Davies-Bouldin index (DB) is the ratio of within-clusters and between-clusters similarities and should be minimized. The silhouette score is calculated using the mean intra-cluster distance and the mean nearest-cluster distance and should be maximized. To compute these indices on mixed data, we have to perform a dimensionality reduction, using FAMD or Laplacian Eigenmaps. Silhouette score can also be computed using a distance suited for mixed data as Gower's distance.

	FAMD			Laplacian Eigenmaps			Gower
Algorithm	CH	DB	Silhouette	CH	DB	Silhouette	Silhouette
K-Prototype	111.12	1.899	0.140	**365.85**	4.628	**0.129**	0.213
Kamila	24.95	4.335	-0.128	36.86	9.650	-0.184	-0.087
FAMD + K-Means	**434.49**	0.707	0.563	101.30	5.137	-0.136	**0.338**
FAMD + Pretopo	362.93	**0.622**	**0.571**	98.66	**4.419**	-0.173	0.303

TAB. 1 – Results on the Palmer Penguins Dataset, with k=12 clusters

Table 1 presents the result on the Palmer Penguins Dataset where the parameter K (K-Prototype and K-means) has been computed with the Elbow method. Our approach seems to provide coherent results compared to the commonly used algorithms. Low Davies-Bouldin indices indicate clearly-separated clusters, both with FAMD and Laplacian Eigenmaps reductions.

4 Conclusion

Mixed data clustering is relevant in several fields such as building performance analysis. In this article, we have demonstrated that dimension reduction combined with pretopological hierarchical clustering was a relevant method of mixed dataset clustering. Experimental results show, with different cluster analysis methods, that it can be comparable to or better than other state-of-the-art algorithms. By working hyperparameter tuning on our approach, we hope to obtain better results.

Références

[1] United Nations Environment Programme (2022). *Global Status Report for Buildings and Construction : Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector*. Nairobi.

[2] Amir Ahmad and Shehroz S Khan. Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access*, 7 :31883–31902, 2019.

[3] Efthymios Costa, Ioanna Papatsouma, and Angelos Markos. Benchmarking distance-based partitioning methods for mixed-type data. *arXiv preprint arXiv :2203.16287*, 2022.

[4] Loup-Noé Lévy, Jérémie Bosom, Guillaume Guerard, Soufian Ben Amor, Marc Bui, and Hai Tran. Hierarchical clustering of complex energy systems using pretopology. In *VEHITS, SMARTGREENS*. Springer, 2022.

[5] Julio-Omar Palacio-Niño and Fernando Berzal. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv :1905.05667*, 2019.