Stage Recherche et Développement ESILV

Titre : Challenges of Mixed Data Clustering

Supervisors: Sonia DJEBALI, Enseignant-Chercheur, DVRC
Guillaume GUERARD, Enseignant-Chercheur, DVRC

Industrial context

The energy sector is in the midst of significant transformation, prompted by the need to increase the use of renewable energy sources and improve energy efficiency, becoming a Smart Grid. This cutting-edge technology allows for the analysis, management, and coordination of energy production, consumption, and distribution, all with the goal of promoting more sustainable practices. A challenge arises from the fact that the data is mixed, containing both numerical and categorical information, often in the form of a data stream. Analyzing this kind of data requires adapted methods. As a result, traditional methods that are designed for numerical data are not well-suited to this type of data.

Advanced tools for analyzing complex systems that can handle rich and heterogeneous data are crucial for Trusted Third Parties for Energy Measurement and Performance to provide independent energy performance analysis and recommendations for clients. It is important that these tools are also easily interpretable by energy experts to facilitate classification and recommendation.

Creating clusters of similar buildings is an effective way to handle complex energy data. Hierarchical clustering of mixed data is a crucial approach that allows energy experts to easily associate clusters with recommendations. It is an essential tool for not only the energy sector but also has diverse applications in fields such as biology, medicine, marketing, and economics.

Scientific context

Although mixed data is widespread, clustering tools specifically designed for it are limited. Some of the bottlenecks have already been defined in a previous scientific paper. Here is a non-exhaustive list of bottlenecks one can encounter when handling mixed data in a pipeline:

- *Data preprocessing:* Data preprocessing is a critical step in mixed data clustering like handling missing data, encoding categorical data, and scaling numerical data.

- *Feature selection:* Mixed data clustering requires feature selection to be performed before clustering. However, selecting relevant features can be a challenging and time-consuming task.
- *Metric selection:* Choosing the right distance metric to measure the similarity between different data types.
- *Evaluation*: There is a lack of standard evaluation criteria for mixed data clustering, which makes it hard to compare different methods.
- *Computational complexity*: Mixed data clustering involves dealing with different types of data and distance metrics, which can result in high computational complexity.
- *Visualization:* It is difficult to create visualizations that effectively communicate the relationships between different data types.
- *Interpretation:* Understanding the relationships between different data types can be challenging, especially if the clusters are not well-separated or the data are altered before using any methods.

References the read to understand the topic

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. Ieee Access, 7, 31883-31902.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, 104743.

Lévy, L. N., Bosom, J., Guerard, G., Amor, S. B., Bui, M., & Tran, H. (2021). Application of Pretopological Hierarchical Clustering for Buildings Portfolio. In SMARTGREENS (pp. 228-235).

Lévy, L. N., Bosom, J., Guerard, G., Amor, S. B., Bui, M., & Tran, H. (2022). DevOps Model Appproach for Monitoring Smart Energy Systems. Energies, 15(15), 5516.

Amor, S. B., Choufa, M., Cornet, C., Djebali, S., Guerard, G., Lévy, L. N., & Tran, H (2023). Pretopology-based Clustering for Mixed Data. In ROADEF2023

Amor, S. B., Choufa, M., Cornet, C., Djebali, S., Guerard, G., Lévy, L. N., & Tran, H (2023). Clustering Mixed Data Comprising Time Series. In SOICT2023

Profil et Compétences attendues

Étudiante ou étudiant de niveau M2 en informatique (Master ou école d'ingénieurs).
Connaissances en python, méthodes non-supervisées, traitement des données mixtes

Lieu du stage

Laboratoire de recherche De Vinci Research Center au sein de l'École Supérieure d'Ingénieurs Léonard de Vinci ; Paris, la Défense.

Période

Stage de 4-5 mois à effectuer à partir de mars - début avril 2023 (900€ pour M2).

Candidature

Les candidat.e.s sont invité.e.s à nous envoyer un mail à sonia.djebali@devinci.fr avec :
CV indiquant leurs expériences et compétences
Une lettre de motivation
Les bulletins de notes des deux dernières années.