Unsupervised Learning - Project 2022/23

Cameron Lorenzo
Univeristy of Milan - Bicocca
l.cameron@campus.unimib.it

Ighissou Ayoube
Univeristy of Milan - Bicocca
a.ighissou@campus.unimib.it

Abstract*— Clustering analysis plays a vital role in uncovering hidden patterns and structures within datasets. However, working with mixed datasets, which consist of a combination of categorical and numerical features, presents unique challenges. In this report, we explore different clustering methods to tackle this task. More in detail, we employ partitioning clustering using kmodes and kprototypes, hierarchical clustering with the Gowder distance, and deep learning-based clustering using self-organizing maps (SOM). We address the optimal number of clusters through techniques such as the elbow method, silhouette score, and dendrograms while the evaluation of the results is performed using, again, the silhouette score and the Adjusted Rand index for multiple target features. Despite the suboptimal solutions, our study provides insights into the clustering of this type of dataset. We discuss the difficulties tied to working with them, the ongoing research in the field, and the need for more advanced techniques that preserve information and address dimensionality challenges. This research contributes to the broader understanding of clustering mixed datasets and sets the stage for future advancements in this area.

1. Introduction

Clustering is a fundamental task in data mining and machine learning that aims to partition a dataset into groups based on the similarity of its objects. It plays a crucial role in various domains, including pattern recognition, image analysis and anomaly detection. Clustering algorithms facilitate data exploration, providing insights into the underlying structure of the data, and supporting decision-making processes.

In this paper, we present a project focused on developing software for performing clustering analysis on a given dataset, exploring different algorithms, comparing their performance, and identifying the optimal number of clusters for the dataset.

The project involves several tasks that are integral to the clustering process. Firstly, we will develop code for preprocessing the dataset, which includes, among other steps, computing the distance between different data objects. It is important to note that the dataset contains binary and discrete attributes, which pose specific challenges during the preprocessing stage.

Next, we will implement two types of clustering algorithms on the dataset. However, we will ignore certain variables, namely Hypertension, Stroke, and Diabetes, during the clustering process. These variables are excluded due to their potential impact on the clustering results, and focusing on the remaining attributes will allow us to analyze the dataset's structure more effectively.

To compare the performance of different clustering algorithms, we will develop software that evaluates various unsupervised and supervised performance measures. Unsupervised performance measures assess the quality of clustering without any external labels, while supervised performance measures utilize the aforementioned excluded variables to evaluate the clustering results. By employing both types of measures, we can gain a comprehensive understanding of the algorithms' effectiveness and identify any discrepancies between their performances.

In addition, we will implement a section for selecting the optimal number of clusters. This process involves employing different relative measures to assess the clustering solutions produced by various algorithms and it is essential to ensure meaningful interpretations of the data and enhance the overall accuracy of clustering results.

Throughout this paper, we will provide detailed explanations of the implemented software codes, discuss the methodology employed, and present the results obtained from the clustering analysis. By addressing these tasks, we aim to provide valuable insights into the strengths and weaknesses of different algorithms when applied to the given data set.

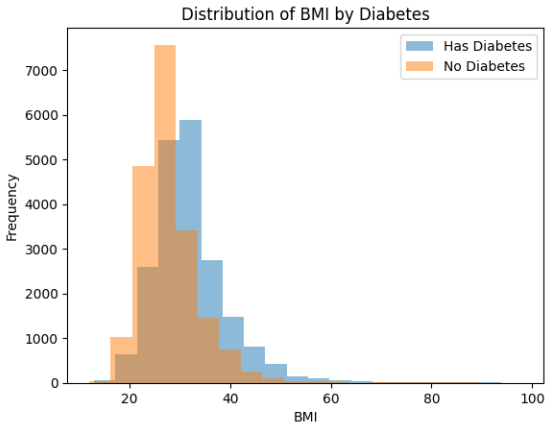
2. Dataset

The original data set consists of 40108 rows and 18 variables. As mentioned in the introduction, the variables are of different types, categorical and numerical. More specifically we can divide them into three subsets:

- Binary variables
- Multilevel categorical variables
- Discrete numerical variables

During an initial exploration stage, we discovered that the number of missing values inside the data set was equal to zero while, on the other hand, we noticed the presence of duplicates (in particular 2456 rows), and we decided to

remove them. After this operation, the new shape of the dataset was now 37652x18. Due to the intrinsic complexity of the dataset, mainly caused by its mixed nature, we spent a consistent amount of time and work during this initial phase of preprocessing and exploration, in order to understand in depth the features, the relationships between them and their hypothetical impact during the later stages of the project. More in detail, our analyses are based on heuristic facts, thus derived primarily from the intrinsic nature of the dataset, and on certified external sources. We would like to emphasize that our understanding and management of the dataset features derive from a preliminary study of the scientific-medical literature. Our focus revolved mainly around two specific variables: Body Mass Index and Age. Regarding the first one, the values within the column range between a minimum of 12 and a maximum of 98. This index generally represents a person's weight in kilograms divided by the square of the height in meters, in other words a simple and inexpensive screening method for identifying the weight categories: underweight, healthy weight, overweight, and obesity. After a further investigation [1], we discovered that this index value range is [19,35]. Once this was established, we tried to furtherly discretize the BMI feature in the dataset, grouping these values in the above-mentioned categories. This was possible also thanks to the correlation between the BMI and a person's age. In our problem dataset, the feature **Age** is based on the **AGE5YR** scale, which represents a breakdown of individuals by age starting from 18 years old. This restriction allowed us to use the mentioned encoding of the BMI variable, since it is aligned with the age requirements. From the graph we can see how the BMI distribution among individuals, with and without diabetes, while below there is the table related to the BMI mapping according to the categories:



Category	Number	Percentage
UnderW	382	1.015
Normal	37236	98.89
OverW	30	0.083
Obesity	4	0.002

From the table clearly emerges that the majority of individuals belong to the **Normal Weight** class. This has the consequence of making the BMI quite non-informative and non-discriminative. To further validate this thesis, we noted that, although excess weight is an important cause of insulin resistance and therefore a significant risk factor for diabetes (in type 2 diabetes in particular), the BMI, despite being correlated with the amount of body fat, does not directly measure its quantity or areas distribution and, therefore, is not always sufficient to carry out an accurate classification of obesity.

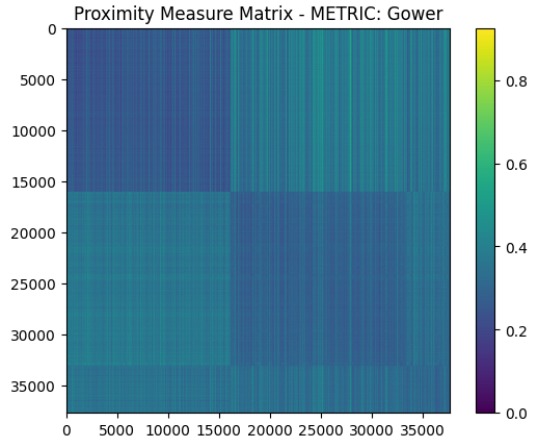


Figure 1: Dissimilarity Matrix

3. Data Exploration - Dissimilarity measure

At this stage of exploring the data set, our goal is to further understand the intrinsic nature of the data, trying to find out if there are correlations between the variables and if clustering patterns can already be identified within the dissimilarity matrix. In order to compute this, we used the so-called Gower distance [2]. The Gower distance can be used to measure how different two records are, and it works well in those domains where the variables are categorical, numerical, logical or even text data. Consequently, being in this situation with both numerical and categorical variables, this measure fits well our purpose. The resulting dissimilarity matrix can be seen in the (fig. 1)

As shown in the matrix, for very low values, i.e. for values tending towards dark blu, we have that the variables in that case are very similar, which therefore leads us to assume initially, without any clustering methods, that the number of clusters within the data set is 3.

4. Models

In this section, we discuss the different clustering methods used to tackle the task, whose choice was guided by the mixed nature of the dataset. It is important to note that clustering mixed datasets is still an active area of research, and various approaches exist, thus our algorithm choices were based on the literature we reviewed, computational considerations, and personal judgment. The decision to implement fundamentally different methods, such

as partitioning clustering, hierarchical clustering, and deep learning-based clustering, was motivated by the expectation that each method may provide unique insights and uncover different structures within the dataset. By approaching the task from multiple perspectives, we aimed to improve the overall understanding of the data and potentially enhance the clustering results.

- **Partitioning Clustering with kmodes and kprototypes:** To handle the mixed nature of the dataset, we employed partitioning clustering algorithms, namely kmodes and kprototypes, both of which are based on the K-means algorithm.

  kmodes is an algorithm specifically designed for clustering categorical data. We utilized kmodes to cluster the categorical features in the dataset, including binary and multi-valued categorical variables. To ensure compatibility, we performed a bin discretization step on the discrete numerical features before applying one-hot encoding. The bin discretization allowed us to convert the numerical values into categorical representations suitable for kmodes. Following this step, we applied one-hot encoding on the discrete numerical features, as well as the multi-valued categorical ones, and combined them with the existing binary variables.

  kprototypes, unlike kmodes, is capable of handling mixed data types, including both categorical and numerical variables. Therefore, there was no need for bin discretization in this case. We simply specified the indices of the categorical features and applied kprototypes directly to the mixed dataset. The algorithm combines K-means for numerical variables and kmodes for categorical variables, resulting in a unified approach to clustering mixed data. It adapts the optimization process to account for the different natures of the data types. It optimizes the cluster centroids by minimizing a combined cost function that considers both the distance between numerical values and the dissimilarity of categorical values.

- **Hierarchical Clustering with Gowder Distance:** To explore alternative clustering methods, we turned to hierarchical clustering, which forms clusters based on the proximity between data points. However, since the dataset contained mixed variables, we employed the Gowder distance to compute the proximity matrix required for hierarchical clustering. The Gowder distance is a dissimilarity measure that accommodates different data types, including categorical (binary and multi-valued) and numerical variables. We utilized it to calculate the dissimilarity between data points with mixed attributes. By employing the Gowder distance, we obtained a proximity matrix that captured the dissimilarity information required for hierarchical clustering.

- **Self-Organizing Maps (SOM) for Deep Learning-based Clustering:** As an additional approach, we ended up considering the SOM [3]. The algorithm uses an unsupervised and iterative procedure to model an input space with a fixed lattice space. In other words, we map all the data points into a lower dimensional space and cluster the points in this new space. The SOM algorithm has several tuning parameters that are set prior the execution and affect the structure of the generated map, as well as the training procedure. Before the training process begins, the weight vectors of the neurons are initialized. This can be done randomly, such is it done in our case, or using some other strategy, for example principal component analysis. In our case, we initialized as 10x10 the dimension of the low dimension space.

5. Optimal Number of Clusters and Evaluation

Determining the optimal number of clusters is a crucial step in clustering analysis. We employed several evaluation techniques to identify the appropriate number of clusters for each clustering method used in our study.

- **Kmodes and Kprototypes:** To determine the optimal number of clusters for the partitioning clustering methods, namely kmodes and kprototypes, we applied the elbow method and silhouette score.
  - **Elbow Method:** The elbow method involved plotting the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS measures the compactness of the clusters and quantifies the variability within each cluster. We looked for the ”elbow” point in the plot, which represents a significant drop in the WCSS. This point indicated the optimal number of clusters where adding more clusters did not lead to a substantial improvement in clustering performance.
  - **Silhouette Score:** The silhouette score is a metric that measures the compactness and separation of clusters. It quantifies how well each data point fits within its assigned cluster compared to other clusters. However, due to the use of a different distance metric, not the Euclidean distance, we had to compute a distance matrix consistently with kmodes and kprototypes. We then calculated the silhouette score using this distance matrix and selected the number of clusters that maximized the average silhouette score, indicating better-defined and well-separated clusters.

  By applying the elbow method (fig. 2 and 4) and silhouette score to kmodes and kprototypes, we were able to determine the optimal number of clusters for each method based on their respective evaluation

metrics, with both method suggesting an optimal number of clusters of 2.

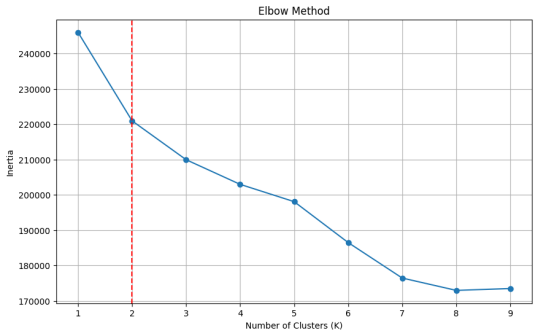


Figure 2: K-Modes elbow method

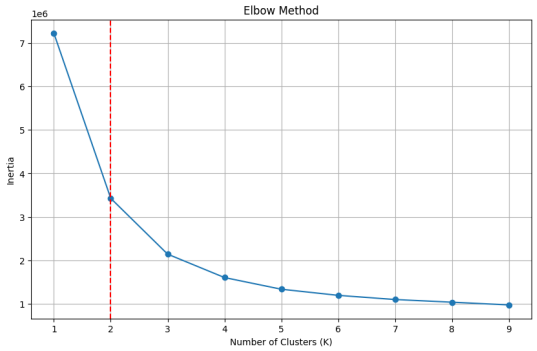


Figure 3: K-Prototype elbow method

- **Hierarchical Clustering:** For hierarchical clustering, we employed a different approach to assess the optimal number of clusters, namely a dendrogram, which visually displayed the merging and splitting of clusters at each level of the hierarchy. A dendrogram is a tree-like diagram that illustrates the hierarchical relationships between clusters. We analyzed the dendrogram to identify the level at which the branches of the tree exhibited a significant increase in length or distance. This length increase indicated a considerable merging of clusters, suggesting a potential optimal number of clusters.

  By examining the dendrogram (fig. **??**), we determined the optimal number of clusters for hierarchical clustering based on the structure and interpretation of the diagram. Additionally, we deemed suitable a double-check with the silhouette score, and both converged again to an optimal number of 2

For evaluation, we chose as unsupervised metric the silhouette score and as supervised the Rand index.

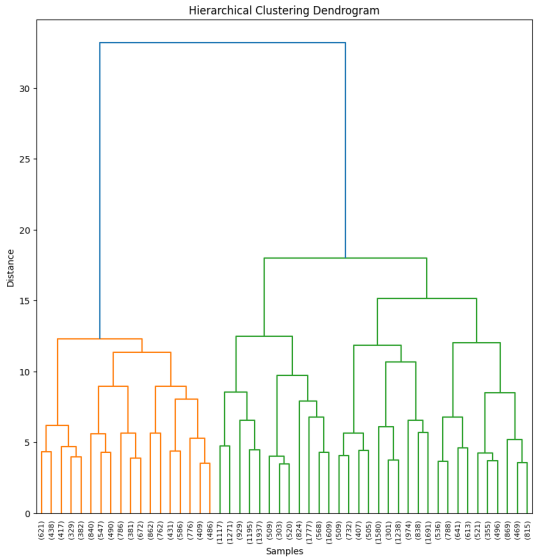


Figure 4: Hierarchical clustering dendogram

- **Silhouette Score:** As mentioned earlier, we used the silhouette score to evaluate the quality of the clustering results. The silhouette score considered the compactness and separation of clusters. Since we used a different distance metric, both for the partitioning and hierarchical methods, we computed the silhouette score based on different distance matrices computed ad hoc, specifically derived for these methods.
- **Adjusted Rand Index:** The adjusted Rand index is a measure of the similarity between two data clusterings, taking into account the expected similarity by chance. We evaluated the adjusted Rand index three times, considering each of the target features (Stroke, Hypertension, and Diabetes) individually. This allowed us to assess the agreement between the clusters generated by the methods and the actual target labels, while considering the possibility of random agreement.

The results obtained are displayed in the table below:

Summary Metrics			
Metrics	KM	KP	HC
Silhouette Score	2.55	0.0988	0.2445
ARI (Stroke)	-0.0031	0.0988	0.0888
ARI (Hypertension)	0.0356	-0.0037	0.0228
ARI (Diabetes)	0.0284	0.0110	0.0515

For what concerns the SOM model [4], its evaluation is mainly based on the measurement quality called quantization error (QE), a measure of the average distance between the data points and the map nodes to which they are mapped, with smaller values indicating better fit. The value for a map is computed using the following formula:

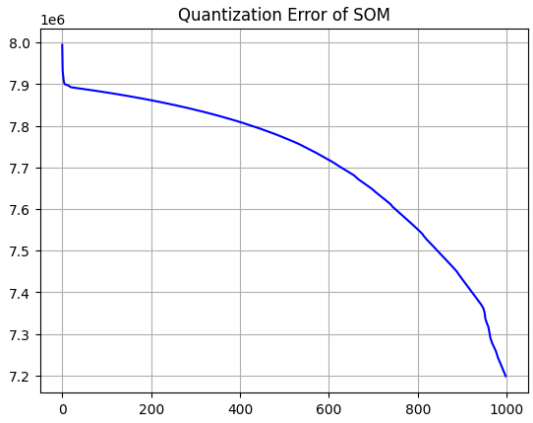


Figure 5: QE Error

$$QE(M) = \frac{1}{n}\sum_{i=1}^{n} \|\phi(x_i) - x_i\|$$

where n is the number of data points in the training data and Φ is the mapping from the space of input data to the lattice of SOM. After a phase of hyperparameter tuning we ended up with the plot of the best quantization error (fig. 5) over a number of training epochs equal to 1000 was in the case of a number of cluster equal to 3.

6. Conclusion

In this study, we addressed the challenging task of clustering a mixed dataset consisting of categorical (binary and multi-valued) features, as well as discrete numerical variables. We employed various clustering methods, including kmodes, kprototypes, hierarchical clustering with the Gowder distance, and deep learning-based clustering using SOM. As previously said, each method was chosen to tackle the task from a different perspective, aiming to gain a comprehensive understanding of the underlying structures within the data.

Working with a mixed dataset presented several difficulties. The heterogeneity of the data types required careful consideration and adaptation of the clustering algorithms. While we applied encoding techniques such as one-hot encoding and bin discretization to handle the mixed features, it inevitably led to information loss. We encountered difficulties in accurately capturing the relationships and patterns in the data, particularly with respect to the ordinal information and the inherent structure of the mixed features. Furthermore, the increased dimensionality resulting from encoding posed challenges in terms of computational efficiency and performance.

The clustering of mixed datasets remains an active area of research. Finding optimal solutions for clustering mixed data is an ongoing challenge. Our choice of clustering methods, as mentioned earlier, was based on the literature, computational considerations, and personal judgment. However, it is important to acknowledge that these methods represent suboptimal solutions due to the inherent limitations and trade-offs associated with handling mixed datasets. Despite these limitations, our approaches provided valuable insights into the data and revealed potential clusters within the mixed dataset.

Moving forward, it is essential to explore more advanced techniques and methodologies that can effectively handle mixed data without compromising important information. Researchers are actively working on developing clustering algorithms that can better accommodate the complexities of mixed datasets, leveraging advancements in machine learning and data mining [5]. These future developments hold promise for improved clustering performance and a deeper understanding of the inherent structures within mixed datasets. [**?**]

References

[1] CDC, "All About Adult BMI — cdc.gov."

[2] D. Anand, "Gower's Distance — medium.com." https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553. [Accessed 21-Jun-2023].

[3] Chung-Chian Hsu, "Generalizing self-organizing map for categorical data."

[4] Gregory T. Breard , "Evaluating self-evaluating self-organizing map quality measure organizing map quality measures as convergence criteria," 2017.

[5] "Survey of State-of-the-Art Mixed Data Clustering Algorithms — iee-explore.ieee.org." https://ieeexplore.ieee.org/document/8662561. [Accessed 21-Jun-2023].

We declare that what has been written in this work has been written by us and that, with the exception of quotations, no part has been copied from scientific publications, the Internet or from research works already presented in the academic field by us or by other students.