Bachelor Thesis Double bachelor BSc^2 in Econometrics and Economics

Erasmus University Rotterdam

Erasmus School of Economics

Enhancing the interpretability of cluster analysis based on cluster explanations that are both accurate and distinctive by enlarging its applicability to mixed-type data

Author:
Fleur van Exel* (544966)

Supervisor:
Rick Willemsen

Second assessor:
Wilco van den Heuvel

Date final version: 30th June 2023

Abstract

Despite the many existent types of clustering algorithms, clustering mixed-type data in an interpretable way remains a challenging problem. In this paper, Mixed Integer Linear Optimization problems are introduced for finding cluster explanations, based on the characteristics of the observations, when cluster allocations are already known or for jointly finding clusters and their explanations. To examine which out of three distance measures is the most effective in assigning observations to clusters and finding the corresponding cluster explanations, their results are compared regarding the accuracy, being the fraction of observations within the cluster satisfying the cluster's explanation, distinctiveness, being the fraction of observations outside the cluster that is true to the cluster's explanation, and cluster explanations. The extended unweighted Gower distance is preferred over the squared Euclidean distance and the unweighted Gower distance since, in case of a different result between the tree distance measures, it either provides better cluster allocations or finds shorter cluster explanations. The results are obtained by analyzing six datasets differing in their feature types.

*The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

1 Introduction

Cluster analysis is the task of dividing observations in a dataset into different groups, based on their characteristics, in such a way that the observations within a group are as similar as possible but as dissimilar as possible to observations outside the group. Cluster analysis arises, among others, in the domain of identifying fake news (Zhang, Gupta, Kauten, Deokar & Qin, 2019), customer segmentation (Mihova & Pavlov, 2018), spam filtering (Nagwani & Sharaff, 2017), detecting fraudulent or criminal activity (Prabakaran & Mitra, 2018), document organization (Cui, Potok & Palathingal, 2005), player classification (Ramirez-Cano, Colton & Baumgarten, 2010), identifying risk factors in healthcare (Kolbe-Alexander, Conradie & Lambert, 2013) and astronomy analysis (Jang & Hendry, 2007). Although the function of state-of-the-art clustering algorithms is to discover and explain groups, they provide little insights (Bertsimas, Orfanoudaki & Wiberg, 2021). Therefore, the goal of this research is to enhance the interpretability of cluster analysis by enlarging its applicability based on cluster explanations that are both accurate and distinctive.

Carrizosa, Kurishchenko, Marín and Romero Morales (2023) introduce a new approach to making cluster analysis more interpretable. Rather than having long and hard-to-grasp explanations for each cluster, the focus is on reducing the complexity to two simple statements while maintaining clustering performance. Their approach is, however, limited to a distance measure that is suitable for continuous features only (Lee, 2022), whereas most real-world related databases consist of mixed-type data types. For example, employees have a gender (categorical), an income (continuous), an educational level (ordinal) and are applicable for bonuses or not (binary). Thereby, it is important to note that binary features are a special case of categorical features having two categories and that ordinal features are categorical features with a clear ordering of the categories (Palmer, 2019). Via the usage of the Gower distance, the applicability of the approach of Carrizosa et al. (2023) can be enlarged by allowing it to account for more types of features, which is also this paper's contribution to the current literature, and therefore to strengthen its interpretability.

Two mathematical optimization problems are introduced, namely one for explaining clusters when cluster allocations are already known and one for finding clusters and their explanations simultaneously. The extended unweighted Gower distance, if not performing similarly to the squared Euclidean distance or unweighted Gower distance, either has the benefit of providing better cluster allocations or finding shorter cluster explanations. These explanations are based on the observation's characteristics combined with the AND operator. To ensure these explanations are easy to understand, the length of the explanation is limited to a single AND-contraction.

The remainder of this paper is organized as follows. Section 2 covers a literature review on cluster interpretability and mixed-type data clustering. Hereafter, Section 3 presents the problem description and Section 4 introduces the mathematical optimization problems that solely provide explanations when clusters are given or simultaneously find clusters and their explanations. Section 4 discusses the various datasets retrieved, followed by the results of the mathematical optimization problems in Section 5. Lastly, some concluding remarks are made in Section 6.

2 Literature review

Two major challenges of cluster analysis are cluster interpretability and mixed-type data (Plant & Böhm, 2011). Plant and Böhm (2011) highlight that most clustering algorithms solely assign observations to a cluster without explaining why or what distinguishes one cluster from another. Moreover, most clustering algorithms are restricted to a single-typed feature only (Foss & Markatou, 2018), and the ones that exist for mixed-type data generally use different optimization goals or separate cluster analysis for continuous and categorical features, sometimes resulting in contradicting conclusions (Hendrickson, 2014).

2.1 Cluster interpretability

With the nature of cluster analysis to distinguish and describe the different clusters obtained, it is usually not sufficient to solely separate observations into the different clusters without explaining the clusters themselves (Chen, 2018). According to Chen (2018), interpretable clustering refers to clustering algorithms that provide cluster explanations and explain the differences between the clusters. Even though the importance of interpretable clustering is well-understood regarding scientific understanding, safety, ethics, mismatched objectives and trade-offs (Doshi-Velez & Kim, 2017), the explanatory power of state-of-the-art clustering algorithms still lacks (Bertsimas et al., 2021). Interpretable clustering can be subdivided into two movements (Lawless, Kalagnanam, Nguyen, Phan & Reddy, 2022). Post-hoc algorithms take the output of a clustering algorithm and try to fit an explanation to the different clusters, whereas integrated interpretability clustering algorithms find clusters and explanations simultaneously.

Although most clustering algorithms focus more on performance than interpretability (Bertsimas et al., 2021), and thus require post-hoc algorithms for interpretability, post-hoc algorithms have the risk of creating explanations that are implied by an algorithm itself rather than inferred from the data (Laugel, Lesot, Marsala, Renard & Detyniecki, 2019). Kenny, Delaney, Greene and Keane (2021) elaborate heron by making the distinction between integrated algorithms that directly explain how a model is optimized using such-and-such methods or post-hoc algorithms that explain or justify the obtained solution by showing how this solution was reached since such-and-such data was used. Integrated interpretable clustering methods, on the other side of the spectrum, mainly cover decision trees, according to Bertsimas et al. (2021), in which features split the observations into different leaves. Clusters are represented by (the conjunction of) leaves, where the path from the root to the leaves (or a single leaf) corresponding to the cluster forms the explanation of that cluster. A cluster spread across different leaves thus combines paths using the OR operator (Carrizosa et al., 2023). Carrizosa et al. (2023) therefore introduced an approach based on linear programming that limits the complexity of the rules assigned to a cluster by restricting it to the AND operator only. The complexity of a single cluster could, for instance, be reduced from ((LSTAT $\leq$ 9.95) AND (RM $>$ 6.12)) OR ((LSTAT $>$ 9.95) AND (TAX $\leq$ 302)) for Classification and Regression Trees to (LSTAT $\leq$ 11.36) AND (RM $>$ 6.086) while keeping similar performance regarding homogeneity, accuracy and distinctiveness.

2.2 Clustering mixed-type data

Even though clustering algorithms for handling a single-typed feature are prevalent, few exist for mixed-type data, and those that exist are often an imperfect utilization of approaches designed for single-typed data (Foss & Markatou, 2018). More specifically, features are commonly transferred to adhere to a specific feature type leading to, among others, a loss of information or an increase in the number of dimensions (Foss & Markatou, 2018). The literature on clustering mixed-type data is subdivisible into three literary movements (Tran, Fan & Shahabi, 2021), namely transforming all features into continuous ones, transforming all features into categorical ones or handling mixed-type data.

2.2.1 Ensure all features are continuous

The first branch of research concerns the conversion of non-continuous features into continuous ones and then applying clustering algorithms designed for continuous features. In the case of solely continuous features, the similarity between observations is more intuitive as, for example, a 55-year-old individual is more similar to a 45-year-old than to a 25-year-old (Lasaosa, 2021). The distinction for non-continuous features is, however, less trivial since the difference between single, married or divorced is not so obvious. Using an inaccurate encoding technique to transform non-continuous features into continuous ones can therefore result in a misunderstanding of the data by the clustering algorithm and thus an ensemble of meaningless clusters (Lasaosa, 2021). Moreover, since basic encoding approaches for converting non-continuous to continuous features operate on single features at the time, the newly created features do not capture the correlation between the different features correctly (Tran et al., 2021). Other disadvantages of encoding are an increased number of dimensions (one-hot encoding) and the imposition of non-existent relationships (ordinal encoding).

One-hot encoding converts a categorical feature having c categories into c dummy features assigning the value one to the binary features corresponding to its value and the value zero to the other binary features. This procedure increases the number of dimensions substantially when the number of categories is large. For example, the dataset on rental listings from the American site Craigslist (van Fraassen, Hensen, de Wind & van Exel, 2023) contains the categorical feature *region* with 404 different categories. One-hot encoding of this feature would highly increase the dimensionality and sparsity since each observation now has an additional 403 features containing the value zero. Moreover, as the number of features increases, the performance of continuous-features-based measures like the Euclidean distance metric and the k-means method worsens, as, respectively, poorer contrast between the furthest and nearest neighbor can be provided (Aggarwal, Hinneburg & Keim, 2001) or the method becomes slower and less effective mainly because the data has become more sparse (Boehmke & Greenwell, 2019).

Ordinal encoding enforces an ordered output that is not necessarily existent in the data. Tran et al. (2021) provide the example of a dataset containing information on the *area* (continuous) and *color* (categorical) of four different *shapes* (categorical). The *shapes* considered are triangles having the color blue, circles having the color red, diamonds having the color red and squares having the color blue. From the data, it could be induced that the blue *shapes* (i.e., the triangles and squares) tend to have a small *area* and therefore triangles are more similar to squares than to circles. However, ordinal encoding (i.e., triangle → 0, circle → 1, diamond → 2 and square → 3) imposes an ordinal relationship, meaning triangles are closer to circles than to diamonds, although this ordering does not hold in reality. Moreover, for ordinal features, measuring distances between the different values might be unsuitable since they need not be equidistant (D'Orazio, 2021). Løvik, Siglen and Bjorvatn (2022) provide the example of asking for customer satisfaction measured on a seven-point scale ranging from *highly dissatisfied* to *highly satisfied*. The options are *highly dissatisfied*, *dissatisfied*, *somewhat dissatisfied*, *neither dissatisfied nor satisfied*, *somewhat satisfied*, *satisfied* and *highly satisfied*. Taking *highly dissatisfied* as equal to one and *highly satisfied* as equal to seven, it is assumed that the distance between *highly dissatisfied* and *dissatisfied* is the same as the distance between *neither dissatisfied nor satisfied* and *somewhat satisfied*, since both differ one in their values, but this might not be the case (Løvik et al., 2022). Alternatively, one-hot encoding these ordinal values loses information about the order. The distance, for the satisfaction feature solely, between *highly dissatisfied* which is represented by (1,0,0,0,0,0,0) and *dissatisfied* which is represented by (0,1,0,0,0,0,0) is then the same as the distance between *highly dissatisfied* and *highly satisfied* which is represented by (0,0,0,0,0,0,1), since now the difference between two unequal and five equal binary features is taken.

2.2.2 Ensure all features are categorical

Another direction in research is that of converting non-categorical features into categorical ones and consequently applying clustering algorithms designed for categorical features. The transformation of continuous into categorical features, however, loses information since continuous features are discretized into bins (Tran et al., 2021). Harrison and Pius (2020) refer to the case of *age* dichotomized into young and old at 42. Then, a baby and a teenager are alike based on their categorical *age*. Especially if the range of the values in a single bin is broad, like the example of *age*, information is thrown away. Moreover, categorical features lack a natural order, have high dimensionality and are limited to a certain number of dimensions since they fail to cluster data in all dimensions (Saxena & Singh, 2016).

2.2.3 Approaches that handle mixed-type data

The third field of clustering mixed-type data is the application of approaches able to handle mixed-type data. Only a few such approaches exist, out of which the Gower distance is the most popular distance measure (D'Orazio, 2021), and is derived from Gower's similarity (Gower, 1971). To calculate the similarity between two observations, the Gower distance is computed as the weighted average of partial distances across the different features, where the computation of the partial distance depends on the specific feature type and ranges between zero and one. Eventually, the overall Gower distance equals a number between zero and one, with zero implying the two observations are equal and one implying the two observations are as different as possible (Lasaosa, 2021). Moreover, missing values are accounted for since the Gower distance disregards features having a missing value for one of the two observations. Obviously, observations containing missing values for all the features should be dropped (D'Orazio, 2021). Gower (1971) restricted the Gower distance to continuous and categorical features, and Podani (1999) extended it to ordinal features.

D'Orazio (2021), however, thinks of the Gower distance as misleading for two reasons. Firstly, there is an unequal contribution of continuous and categorical features to the overall Gower distance. Since the partial distance between two observations for continuous features is calculated as the absolute difference between the two values divided by the feature's range, it is highly affected by outliers. The partial dependence of a continuous feature thus only reaches a value of zero when the two observations are exactly the same and a value of one when the two observations are on opposite sides of the spectrum. Otherwise, the value is somewhere in between zero and one. For categorical features, the partial distance between two observations equals zero if the two values are the same and one if different. Categorical values thus always equal the minimum or maximum difference possible, whereas this is seldom the case for continuous features. Secondly, this unbalanced contribution of continuous and categorical features becomes more apparent when the number of categorical features increases. For instance, considering *age* as a continuous feature and *gender* and *marital status* as dummy features, two individuals sharing the same *gender* and *marital status* but having a very different age are considered closer to each other than two individuals differing slightly in *age* but having a different *gender* or *marital status* (D'Orazio, 2021). Expanding the number of categorical features by, for example, adding a dummy for whether one wears glasses mitigates the effect of the (difference in the value of the) continuous features even further.

To tackle the first drawback, D'Orazio (2021) suggests dividing the absolute difference in feature values by the 25%-75% inter-quartile range rather than the whole range, for continuous features, permitting less dependency on outliers. To deal with the unbalanced contribution of continuous and categorical features, D'Orazio (2021) sets the distance between observations considered close equal to zero. Observations are considered close when the absolute difference is less than a certain width away based on the kernel density estimation or if the observations are nearest neighbors from each other.

2.3 Contribution

This paper contributes to the current literature by combining the interpretability-based approach of Carrizosa et al. (2023) and the more widely applicable Gower distance (Gower, 1971) extended by Podani (1999) and D'Orazio (2021). Essentially, the method of Carrizosa et al. (2023) is altered to measure distances between observations differently, avoiding the drawbacks of transforming feature types, and extending the scope of features with ordinal features.

3 Problem description

Two different scenarios are considered, namely finding explanations when clusters are already given, referred to as InterP, or simultaneously finding clusters and their explanations, referred to as CinterP. Given the set of observations I, with indices i and j with $i \neq j$, and the non-overlapping clusters K, with index k, the objective of InterP is to find cluster explanations based on the observations' characteristics. These characteristics conform to the different features f which together form the set F. A set of rules N, with index n, functions as candidate splits for the features. For CinterP, on the other hand, only the set of observations is given such that the aim is to find the optimal partitioning of the observations into $|K|$ distinct clusters based on their characteristics.

4 Methodology

This section introduces Mixed Integer Linear Optimization problems (MILPs), as described in Carrizosa et al. (2023), for InterP and CinterP. Hereafter, the different distance measures, the construction of the set of rules and the time complexity are elaborated on.

4.1 Cluster evaluation

For quantifying the clusters' explanations, three criteria are considered (Carrizosa et al., 2023). Firstly, the observations within a single cluster must be as similar as possible, which corresponds to minimizing the cluster's intra-homogeneity that is quantified by a distance measure. Secondly, the cluster's explanation should be true for as many observations within the cluster as possible, referred to as accuracy. The accuracy equals the number of observations within the cluster satisfying the cluster's explanation divided by the total number of observations within the cluster, and thus ideally equals one. Thirdly, the cluster's explanation should be false for as many observations outside the cluster as possible, referred to as distinctiveness. The distinctiveness equals the total number of observations outside the cluster satisfying the cluster's explanation divided by the total number of observations outside the cluster, and thus ideally equals zero. Each observation can only belong to one cluster, meaning there is no overlap between the clusters.

4.2 InterP

To provide explanations when clusters are already known, the accuracy and distinctiveness are optimized. Observations can either be labeled or clusters can be obtained via another clustering algorithm, using this formulation as a post-hoc algorithm.

4.2.1 Mathematical optimization model

Each cluster k is associated with a set of observations g_k such that $\sum_{k=1}^{|K|} g_k = I$ and $g_k \cap g_{k'} = \emptyset$, for $k \neq k'$. The set of rules, which is elaborated on in section 4.5, can be split into S different groups of features, with index s, corresponding to (groups of) the different features, such that $N = \cup_{s=1}^{|S|} N_s$ and $N_s \cap N_{s'} = \emptyset$ for $s \neq s'$. The explanation of a cluster eventually equals the conjunction of rules selected for that cluster, joined by the AND operator. Additionally, the parameter z_{ksn} equals one if, for (the group of) feature(s) s, rule $n \in N_s$ is chosen for the explanation of cluster k, and zero otherwise, with a maximum of λ explanations per cluster joined by the AND operator. Similarly, the parameter b_{isn} equals one if observation i satisfies the explanation $n \in N_s$ and zero if not. All parameters b_{isn} together form the compatibility set B. The decision variable γ_{ki} equals to one if observation i is true to the explanation assigned to cluster k, irrespective of what cluster observation i is in, and otherwise equals zero. The mathematical optimization model can then be formulated as follows:

$$
\begin{aligned}
\min \quad & -\sum_{k=1}^{|K|} \sum_{i \in g_k} \gamma_{ki} + \theta \sum_{k=1}^{|K|} \sum_{k'=1, k' \neq k}^{|K|} \sum_{i \in g_{k'}} \gamma_{ki}, && && (1)\\
\text{s.t.} \quad & \sum_{n \in N_s} z_{ksn} \leq 1, && \forall k \in K, \forall s \in S, && (2)\\
& 1 \leq \sum_{s=1}^{|S|} \sum_{n \in N_s} z_{ksn} \leq \lambda, && \forall k \in K, && (3)\\
& \gamma_{ki} + \sum_{n \in N_s} (1 - b_{isn}) z_{ksn} \leq 1, && \forall k \in K, \forall s \in S, \forall i \in g_k, && (4)\\
& \gamma_{ki} + \sum_{s=1}^{|S|} \sum_{n \in N_s} (1 - b_{isn}) z_{ksn} \geq 1, && \forall k \in K, \forall k' \in K, k \neq k', \forall i \in g_{k'}, && (5)\\
& z_{ksn} \in \{0, 1\}, && \forall k \in K, \forall s \in S, \forall n \in N_s, && (6)\\
& \gamma_{ki} \in \{0, 1\}, && \forall i \in I, \forall k \in K. && (7)
\end{aligned}
$$

The objective (1), respectively, maximizes the number of observations that are true to the explanation of the cluster it is assigned to and minimizes the number of observations outside the cluster satisfying the cluster's explanation weighted by θ. Equal importance of accuracy and distinctiveness implies $\theta = 1$. Constraint (2) ensures only one rule can be chosen for each group of features, together with constraint (6). A minimum of one and a maximum of λ rules can be chosen for each cluster because of constraint (3). Constraints (4)-(5) are introduced to ensure γ_{ki} is well-defined. In light of the direction of the objective function, the definitions of the observations that are false to the explanation of its cluster (corresponding to the first part of the objective function) and observations that are true to explanations of other clusters (corresponding to the second part of the objective function) need to be precise as the objective would otherwise set them equal to one and zero, respectively. In the case of observation i' belonging to cluster k', but not being a true positive case for the cluster k', $\gamma_{k'i'}$ must equal zero. If rule $n' \in N_s$ is chosen, $z_{k'sn'} = 1$ and $b_{i'sn'} = 0$ since observation i' is not explained by the cluster, implying constraint (4) reduces to $\gamma_{k'i'} + 1 \leq 1$. Together with constraint (7), $\gamma_{k'i'}$ then thus equals zero for cluster k'. On the other hand, for all clusters $k \neq k'$, the constraint becomes redundant since $\sum_{n \in N_s}(1 - b_{i'sn})z_{ksn} \leq 1$. If observation i' is also true to the explanation assigned to another cluster, say cluster k'' with $k'' \neq k'$, $\gamma_{k''i'}$ must equal one. Since observation i' satisfies the explanation of cluster k'', $\sum_{s=1}^{|S|} \sum_{n \in N_s}(1 - b_{i'sn})z_{k''sn} = 0$, meaning $\gamma_{k''i'} = 1$, due to constraint (5). Constraint (5) becomes redundant when observation i' is false for the explanation of another cluster.

4.3 CinterP

For the multi-objective MILP formulation that simultaneously finds $|K|$ clusters and their explanations, the intra-homogeneity, accuracy and distinctiveness are optimized jointly.

4.3.1 Mathematical optimization model

The parameter δ_{ij} denotes the distance measure used to quantify the difference between the normalized observations i and j, for which different measures are taken as further explained in section 4.4. Moreover, there are four binary decisions, associating true with one and false with zero, being whether observation i is a member of cluster k, represented by x_{ki}, whether rule $n \in N_s$ is chosen for explaining cluster k, represented by z_{ksn}, whether observation i is true to the explanation assigned to its cluster, represented by a_i, and whether observation i is outside cluster k but true to the explanation assigned to cluster k, represented by β_{ki}. Assuming the number of clusters is known a priori, or otherwise heuristics like the elbow method can be used to determine $|K|$, the formulation of the mathematical optimization model is as follows:

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{|K|} \sum_{i=1}^{|I|-1} \sum_{j=i+1}^{|I|} \delta_{ij} x_{ki} x_{kj} - \theta_1 \sum_{i=1}^{|I|} \alpha_i + \theta_2 \sum_{k=1}^{|K|} \sum_{i=1}^{|I|} \beta_{ki}, && && (8)\\
\text{s.t.} \quad & \sum_{k=1}^{|K|} x_{ki} = 1, && \forall i \in I, && (9)\\
& \sum_{n \in N_s} z_{ksn} \leq 1, && \forall k \in K, \forall s \in S, && (10)\\
& 1 \leq \sum_{s=1}^{|S|} \sum_{n \in N_s} z_{ksn} \leq \lambda, && \forall k \in K, && (11)\\
& \alpha_i + x_{ki} + \sum_{n \in N_s} (1 - b_{isn}) z_{ksn} \leq 2, && \forall k \in K, \forall s \in S, \forall i \in I, && (12)\\
& \beta_{ki} + x_{ki} + \sum_{s=1}^{|S|} \sum_{n \in N_s} (1 - b_{isn}) z_{ksn} \geq 1, && \forall k \in K, \forall i \in I, && (13)\\
& x_{ki} \in \{0, 1\}, && \forall k \in K, \forall i \in I, && (14)\\
& z_{ksn} \in \{0, 1\}, && \forall k \in K, \forall s \in S, \forall n \in N_s, && (15)\\
& \alpha_i \in \{0, 1\}, && \forall i \in I, && (16)\\
& \beta_{ki} \in \{0, 1\}, && \forall k \in K, \forall i \in I. && (17)
\end{aligned}
$$

The objective function (8) covers all three criteria, as described in section 4.1, by minimizing the distance between observations within the same cluster (i.e., the intra-homogeneity), maximizing the true positive rates (i.e., the accuracy) weighted by θ_1 and minimizing the false positive rates (i.e., the distinctiveness) weighted by θ_2. Given each observation is associated with one cluster only, constraint (9) is introduced to ensure this, together with the binary constraint of x_{ki} (14). Moreover, for each group of features, only one rule can be chosen, which is imposed by constraints (10) and (15). For interpretability purposes, a minimum of one and a maximum of λ rules can be chosen for each cluster because of constraints (11) and (15). Constraints (16) and (17) enforce the domain of α_i and β_{ki}, respectively. Without introducing any additional constraints, the objective would set all α_i equal to one and all β_{ki} to zero. Therefore, to ensure these two variables are well-defined, constraints (12) and (13) are added. In the case of observation i' belonging to cluster k', meaning $x_{k'i'} = 1$, but not being a true positive case for the cluster k', $\alpha_{i'}$ must equal zero. If rule $n' \in N_s$ is chosen, $z_{k'sn'} = 1$ and $b_{i'sn'} = 0$ since observation i' is not explained by the cluster, implying (12) reduces to $\alpha_{i'} + 1 + 1 \leq 2$ for cluster k' and thus $\alpha_{i'} \leq 0$. Together with constraint (16), $\alpha_{i'}$ then equals zero for cluster k'. On the other hand, for all clusters $k \neq k'$, the constraint becomes redundant since $x_{ki'} = 0$ and $\sum_{n \in N_s} (1 - b_{i'sn}) z_{ksn} \leq 1$. For $\beta_{k'i'}$, constraint (13) becomes redundant when cluster k' is considered as then $x_{k'i'} = 1$. For any cluster $k \neq k'$, meaning $x_{ki'} = 0$, such that observation i' is true to the explanation of that cluster, both $b_{i'sn}$ and z_{ksn} equal one such that $\beta_{ki'} \geq 1$. Since $\beta_{ki'}$ cannot be larger than one, due to constraint (17), $\beta_{ki'}$ must equal one.

Although the objective function (8) is straightforward to grasp, the mathematical optimization model is non-linear due to the product of the two binary decision variables x_{ki} and x_{kj}. To benefit from the properties of a linear mathematical programming model, which is being faster at achieving the deterministic global optimum (Toragay, 2019), the binary decision variable y_{kij} is introduced and equals to the product of x_{ki} and x_{kj}, meaning $y_{kij} = x_{ki}x_{kj}$. Consequently, the intra-homogeneity part of the objective function is replaced by the linear formulation of $\sum_{k=1}^{|K|} \sum_{i=1}^{|I|-1} \sum_{j=i+1}^{|I|} \delta_{ij} y_{kij}$. To impose y_{kij} is well-defined and only equals one if both x_{ki} and x_{kj} are equal to one, and otherwise becomes redundant, the constraint $x_{ki} + x_{kj} - y_{kij} \leq 1, \forall i \in \{1, ..., I-1\}, \forall j \in \{i+1, ..., I\}, \forall k \in K$ is introduced. With this linearization, the problem can be written as a MILP as follows:

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{|K|} \sum_{i=1}^{|I|-1} \sum_{j=i+1}^{|I|} \delta_{ij} y_{kij} - \theta_1 \sum_{i=1}^{|I|} \alpha_i + \theta_2 \sum_{k=1}^{|K|} \sum_{i=1}^{|I|} \beta_{ki}, && (18)\\
\text{s.t.} \quad & (9) - (17), && (19)\\
& x_{ki} + x_{kj} - y_{kij} \leq 1, \qquad \forall i \in \{1, ..., I-1\}, \forall j \in \{i+1, ..., I\}, \forall k \in K, && (20)\\
& y_{kij} \in \{0, 1\}, \qquad \forall i \in \{1, ..., I-1\}, \forall j \in \{i+1, ..., I\}, \forall k \in K. && (21)
\end{aligned}
$$

To give the three objectives approximately the same scaling, the intra-homogeneity is divided by $I^2 \max_{ij}\{\delta_{ij}^2\}$, the accuracy by I and the distinctiveness by I (Carrizosa et al., 2023).

4.4 Distance measures

To evaluate the effect of the distance measure and thus the quantification of similarity between observations, three different distance measures are compared, which are the squared Euclidean distance, the unweighted Gower distance and the extended unweighted Gower distance. Note that the distance measure only applies to the MILP of CinterP.

4.4.1 Squared Euclidean distance

Due to its convenient properties, among which are symmetry and definiteness, forms of the Euclidean metric are often used as a distance measure in cluster analysis (Mimmack, Mason & Galpin, 2001). The squared Euclidean distance between two observations i and j equals the sum of the squared differences between the corresponding feature values v_i and v_j of, respectively, observations i and j. Categorical features, including binary and ordinal features, are one-hot encoded. The squared Euclidean distance δ_{ij}^{SE} can then be formalized as follows:

$$\delta_{ij}^{SE} = \sum_{f=1}^{|F|} (v_i^{(f)} - v_j^{(f)})^2. \tag{22}$$

4.4.2 Unweighted Gower distance

To calculate the distance between observations i and j, the Gower distance δ_{ij}^{G} is defined as the weighted sum of partial distances, which is the product of the indicator variable $d_{ij}^{(f)}$ and distance variable $\delta_{ij}^{(f)}$, across the different features f, where the computation of the partial distance depends on the specific feature type. Only if the values $v_i^{(f)}$ and $v_j^{(f)}$ are non-missing for observations i and j and feature f, $d_{ij}^{(f)}$ equals one. Otherwise, $d_{ij}^{(f)}$ equals zero, implying $v_i^{(f)}$ or $v_j^{(f)}$, or both, are missing. In other words, $d_{ij}^{(f)} = 0$ implies feature f does not contribute to the overall distance between observations i and j. Because of its nature, indicator variable $d_{ij}^{(f)}$ thus accounts for missing values. The weight $w^{(f)}$ of a single feature towards the overall distance is predefined. Each partial distance $d_{ij}^{(f)}\delta_{ij}^{(f)}$ ranges between (and including) zero and one such that the overall Gower distance also ranges between (and including) zero and one, with zero implying the two observations are equal and one implying the two observations are as different as possible (Lasaosa, 2021). The formula for the Gower distance is as follows:

$$\delta_{ij}^{G} = \frac{\sum_{f=1}^{F} w^{(f)} d_{ij}^{(f)} \delta_{ij}^{(f)}}{\sum_{f=1}^{F} w^{(f)} d_{ij}^{(f)}}. \tag{23}$$

The unweighted Gower distance corresponds to setting the weight of each feature towards the overall distance equal to one, i.e. $w^{(f)} = 1, \forall f \in F$. The unweighted Gower distance is also considered throughout this paper as the weighting scheme is outside the scope of this paper. The unweighted Gower distance then reduces to:

$$\delta_{ij}^{G} = \frac{\sum_{f=1}^{F} d_{ij}^{(f)} \delta_{ij}^{(f)}}{\sum_{f=1}^{F} d_{ij}^{(f)}}. \tag{24}$$

The calculation of $\delta_{ij}^{(f)}$ depends on the feature type of f and its calculation is represented in Table 1. If all features are continuous, binary or categorical, the unweighted Gower distance is equivalent to the Manhattan distance, Jaccard coefficient or Simple Matching Coefficient, respectively (Hajnal & Loosveldt, 1998).

Table 1: Gower distance measure per feature type and type of Gower distance

Gower distance	**Feature type**	$\delta_{ij}^{(f)}$	
Unweighted	Continuous	$\frac{\left\lvert \upsilon_i^{(f)} - \upsilon_j^{(f)} \right\rvert}{max\{\upsilon^{(f)}\} - min\{\upsilon^{(f)}\}}$	(eq1)
	Categorical	$\begin{cases} 1 & \text{if } \upsilon_i^{(f)} = \upsilon_j^{(f)} \\ 0 & \text{if } \upsilon_i^{(f)} \neq \upsilon_j^{(f)} \end{cases}$	(eq2)
Extended unweighted	Continuous	$\begin{cases} 0 & \text{if } \upsilon_j \in V_{\upsilon_i} \\ \frac{\left\lvert \upsilon_i^{(f)} - \upsilon_j^{(f)} \right\rvert}{IQR_f} & \text{if } \upsilon_j \notin V_{\upsilon_i} \text{ AND } \left\lvert \upsilon_i^{(f)} - \upsilon_j^{(f)} \right\rvert < IQR_f \\ 1 & \text{if } \left\lvert \upsilon_i^{(f)} - \upsilon_j^{(f)} \right\rvert \geq IQR_f \end{cases}$	(eq3)
	Categorical	$\begin{cases} 1 & \text{if } \upsilon_i^{(f)} = \upsilon_j^{(f)} \\ 0 & \text{if } \upsilon_i^{(f)} \neq \upsilon_i^{(f)} \end{cases}$	(eq4)
	Ordinal	use (eq3) given $\upsilon_i^{(f)} = \frac{r_i^{(f)}-1}{R^{(f)}-1}$ AND $\upsilon_j^{(f)} = \frac{r_j^{(f)}-1}{R^{(f)}-1}$	(eq5)

Note. $\upsilon_i^{(f)}$ ($\upsilon_j^{(f)}$) represents the value of observation i (j) for feature f, $max\{\upsilon^{(f)}\}$ ($min\{\upsilon^{(f)}\}$) the maximum (minimum) value of feature f over all observations, V_{υ_i} the set of q nearest neighbors of observation i, IQR_f the interquartile range for feature f, $r_i^{(f)}$ the rank number of observation i for feature f and $R^{(f)}$ the maximum rank number of feature f.

4.4.3 Extended unweighted Gower distance

The extended unweighted Gower distance differs in two aspects from the unweighted Gower distance as proposed by Gower (1971). The first extension lies in the alteration of the distance measure for continuous features such that it accounts for the unequal contribution of continuous and categorical features (D'Orazio, 2021). If observation j is one of the q nearest neighbors of observation i, represented by the set V_{υ_i} with $|V_{\upsilon_i}| = q$, the distance between the two observations is set to zero. The value of q is generally set to $\sqrt{n}$ (Arat, 2019), thus this value is also considered throughout this paper. If observation j is not one of the q nearest neighbors of observation i and the absolute difference is larger than the 75% interquartile range, the distance is set to one. Otherwise, the distance is calculated as the absolute difference in feature values of $\upsilon_i^{(f)}$ and $\upsilon_j^{(f)}$ divided by the interquartile range. The second extension lies in the additional distance measure for ordinal features, as proposed by Podani (1999). An ordinal feature f has a given rank, i.e. observation i has rank $r_i^{(f)}$, that is converted to a continuous value based on its rank, after which it is treated like other continuous features. See also Table 1.

4.5 Set of rules

The set of rules N functions as possible candidate splits for the features in explaining the clusters. For example, explaining a cluster of the housing dataset could entail all observations with $CRIM \leq 10$ and $RM > 3$, of which $CRIM \leq 10$ and $RM > 3$ are thus part of the set of rules N. The size of this set strongly depends on the degree of specificity chosen. For continuous features, for instance, all distinct values or percentile thresholds can be taken as possible candidate splits. Considering all the distinct values leads to redundancy since some values are very close to each other and thus yield the same measure of accuracy and distinctiveness (Carrizosa et al., 2023). Using deciles as percentile thresholds not only reduces the number of rules to be considered but also allows for simple interpretation since all observations explained are a multitude of 10%.

For a single continuous or ordinal feature, the form of the rules is $feature_s \leq threshold$ and $feature_s > threshold$ in which the threshold corresponds to the different deciles. If two deciles correspond to the same threshold value, one of them is disregarded to circumvent duplicates. For a categorical feature, the group of rules equals $feature_s = c_1$, $feature_s = c_2$, ..., $feature_s = c_h$ in which c_1, c_2, ..., c_h equal all distinct categories. In the case of a binary feature, this thus reduces to $feature_s = 0$ and $feature_s = 1$. Taking the group of rules for all different groups together forms the set of rules.

4.6 MIP start

To mitigate the effect of the time complexity, a MIP start is added. In essence, a MIP start enables the provision of a starting solution to the solver, like a hint, and therefore tries to speed up the process (IBM, 2022). The MIP start might be a feasible solution to the model as well as an infeasible or even incomplete one. If the provided MIP start is feasible, the input solution serves as an incumbent solution and as a bound for the branch-and-bound algorithm since it eliminates a part of the search space with less-optimal objective values (openletter.mousetail.nl, 2019). If the provided MIP start is infeasible, the solver tries to repair it into a feasible one. If the provided MIP start is incomplete, the solver tries to fill in the missing values in a way it results in a feasible solution.

In this case, the initial solution provided for each run concerns values for the cluster assignments x_{ki} and rules selected z_{ksn}. The k-means algorithm, in which k equals to the corresponding number of clusters the dataset has to be divided into, provides the initial cluster assignments, which are then used in InterP with $\theta = \frac{\theta_2}{\theta_1}$ to obtain the corresponding initial explanations. To be able to replicate the results, a seed is set, which also allows the provision of the same MIP start to the MILPs with different distance measures. The MILPs with different distance measures thus obtain the same initial solution.

5 Data

To evaluate the three different distance measures, four datasets of Carrizosa et al. (2023) are used since they allow for the use of the squared Euclidean distance that is able to handle continuous features only (Lee, 2022). These four datasets entail the *wine*, *glass*, *housing* and *abalone* datasets. The binary feature of the *housing* dataset and the categorical feature of the *abalone* dataset are dummy encoded when the squared Euclidean distance is used as the distance metric. The *abalone* dataset is obtained by drawing a random subset of 835 observations, without replacement, from the original dataset having more than 4000 observations (Carrizosa et al., 2023),

Hereafter, to delve into the shortcomings of the squared Euclidean distance metric, the *contraceptive* and *titanic* datasets are introduced. All six datasets are originally from the UCI Repository (Dua & Graff, 2017) and their characteristics are displayed in Table 2. Although the datasets are originally meant for Supervised Classification, meaning the observations are already labeled to a cluster, CinterP ignores this information and thus tries to assign the observations and interpret the clusters. The description of the datasets with their features, feature description and feature type can be found in Tables A1-A6. Although there are multiple differences between the datasets, the most interesting one is the differences in feature types. Since a distance measure should accommodate the feature types (Petchey & Gaston, 2007) and since the Euclidean distance metric is suitable for continuous features only (Lee, 2022), it is expected that the difference in performance becomes more apparent in the datasets having more features than only the continuous one. It does, however, not imply no differences are expected in the *wine* and *glass* dataset. One of the objectives is namely to minimize the intra-homogeneity, which is achieved by minimizing the distance between the observations within the same cluster, but the distances are measured differently across the three different distance measures.

Table 2: Description of the datasets

Name of dataset	#observations (I)	#features (F)	Feature types	#Classes
wine	178	13	Continuous	3
glass	214	9	Continuous	6
housing	506	13	Continuous, dummy	2
abalone	835	8	Categorical, continuous	2
contraceptive	1473	9	Continuous, dummy, ordinal	3
titanic	712	7	Categorical, continuous, dummy, ordinal	2

Note. Features are displayed to their most specific type, i.e. binary or ordinal is more specific than categorical.

Since the Euclidean distance metric is unable to handle missing values, and to keep other effects between the different models (varying in their distance measure) at a minimum, observations containing missing values are deleted. This results in a deletion of 20.2% of the observations for the *titanic* dataset. The other datasets did not contain any missing values. It is important to note that the deletion of observations with missing values mitigates the ability of the Gower distance to account for them, and thus a benefit of the (extended) unweighted Gower distance over the squared Euclidean distance as a distance measure.

6 Results

This section presents the results of InterP and CinterP. First, some assumptions are made, after which the construction of the rules and the compatibility set is elaborated upon. Hereafter, to verify the implementations of InterP and CinterP, their results are compared to the results of Carrizosa et al. (2023). Afterward, the different distance measures are compared.

6.1 Pre-processing

The MILPs are solved in Java using CPLEX (Cplex, 2009) on a PC Intel(R) Core(TM) i5-8265U with 8.00 GB of RAM. Additionally, a time limit of ten minutes is imposed for a good trade-off between intra-homogeneity, accuracy and distinctiveness (Carrizosa et al., 2023). The number of rules per cluster is limited to two, i.e. $\lambda = 2$, to keep interpretability limited to a single, and thus simple, AND-contraction. Each feature is considered as its own group, implying $S = F$, such that each combination of features can be used to characterize the clusters, with a maximum of $\lambda = 2$. The values considered for θ_1 and θ_2 are combinations of the values 0.5, 1 and 2 such that the accuracy and distinctiveness differ in their relative importance. The values considered for θ are 0.25, 0.5, 1, 2 and 4, which is similar to the fraction $\theta = \frac{\theta_2}{\theta_1}$ that is used for the MIP start. Lastly, normalization is based on the Z-score normalization as this is proposed as a powerful method of normalization for clustering (Mohamad & Usman, 2013).

6.2 Construction of the set of rules and the compatibility set

Before anything can be said about which observations should be allocated to which clusters and which rules should be chosen to explain the clusters, the set of rules must be constructed to know which observations are explained by which rule. This latter information is eventually captured in the compatibility set. The reason these constructions are elaborated upon is that the size of the set of rules differs from that of Carrizosa et al. (2023). The *titanic* dataset is taken as an example since this dataset has the most variety in the type of features. Considering the most specific feature types, the dataset covers the continuous features *AGE*, *NSIB*, *NPAT* and *FARE*, the dummy feature *SEX*, the categorical feature *EMB*, and the ordinal feature *CLASS*. Each feature corresponds to its own group of rules.

6.2.1 Construction of the set of rules

In essence, the set of rules is the same for the squared Euclidean and unweighted Gower distance metric, since both do not distinguish between a categorical and a more specific ordinal feature. In other words, an ordinal feature is treated like the other categorical features. The group of rules considered for categorical features is equal to the distinct categories. The *SEX*, *EMB* and *CLASS* features respectively take the following values $\{0, 1\}$, $\{0, 1, 2\}$, $\{1, 2, 3\}$, resulting in the group of rules as displayed in Table 3. For continuous features, the rules consider $feature_s \leq threshold$ and $feature_s > threshold$ in which the threshold corresponds to the (nine) deciles. Duplicate decile values are disregarded as this leads to redundancy. The decile values for the AGE are $(14.0, 19.0, 22.0, 25.0, 28.0, 31.0, 36.0, 41.0, 50.0)$, for *NSIB* are $(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0)$, for $NPAT$ are $\{0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 2.0\}$, and for *FARE* are $(7.75, 7.8958, 9.0, 12.875, 15.7417, 26.0, 29.0, 46.9, 79.2)$. All together, this thus results in the set of rules as displayed in Table 3.

Table 3: The set of rules for the squared Euclidean and unweighted Gower distance metric for the *titanic* dataset

Feature	Group of rules	Number of rules
SEX	$SEX = 0$, $SEX = 1$	2
EMB	$EMB = 0$, $EMB = 1$, $EMB = 2$	3
CLASS	$CLASS = 1$, $CLASS = 2$, $CLASS = 3$	3
AGE	$AGE \leq 14.0$, $AGE > 14.0$, $AGE \leq 19.0$, $AGE > 19.0$, ..., $AGE \leq 50.0$, $AGE > 50.0$	18
NSIB	$NSIB \leq 0$, $NSIB > 0$, $NSIB \leq 1$, $NSIB > 1$	4
NPA T	$NPAT \leq 0$, $NPAT > 0$, $NPAT \leq 1$, $NPAT > 1$, $NPAT \leq 2$, $NPAT > 2$	6
FARE	$FARE \leq 7.75$, $FARE > 7.75$, ..., $FARE \leq 79.2$, $FARE > 79.2$	18

6.2.2 Extended unweighted Gower

The difference between the squared Euclidean and unweighted Gower distance measure on the one hand and the extended unweighted Gower on the other hand is, concerning the set of rules, the recognition of ordinal features. Therefore, for the extended unweighted Gower distance, the group of rules concerned with the ordinal feature *CLASS* changes into $CLASS \leq 1$, $CLASS > 1$, $CLASS \leq 2$, $CLASS > 2$, $CLASS \leq 3$, $CLASS > 3$, increasing the size of the set of rules by three. Although there are no observations satisfying $CLASS > 3$, this rule should become redundant as none of the observations is true to it. An overview of the total number of rules per dataset and distance measure is provided in Table 4. The relatively large difference in the size of the set of rules of the *abalone* dataset compared to the results of Carrizosa et al. (2023) can be explained by the differences in the subset taken, although the size of the subset is equal. In this case, the continuous features of the *abalone* dataset had twelve duplicate values together, implying 24 rules were disregarded because of redundancy.

Table 4: Overview of the number of rules per dataset and per distance measure

Dataset	squared Euclidean or unweighted Gower distance	extended unweighted Gower distance
wine	234 (235)	234
glass	138 (139)	138
housing	190 (187)	190
abalone	105 (130)	105
contraceptive	52	62
titanic	54	57

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

6.2.3 Construction of the compatibility set

After the set of rules is constructed, it can be derived for all observations by which rules they are explained. Considering two observations of the *titanic* dataset, which are shown in Table 5, the value of b_{isn} can be derived per observation and per feature. Taking the squared Euclidean distance into consideration as the distance metric, the associated set of rules for the *titanic* dataset can be found in Table 3.

Table 5: Two observations from the *titanic* dataset

CLASS	SEX	AGE	NSIB	NPAT	FARE	EMB
3	0	22	1	0	7.25	2
1	1	38	1	0	71.2833	0

Considering the ordinal feature *CLASS*, the first observation is only explained by $CLASS = 3$, implying $b_{1(CLASS)1} = 0$, $b_{1(CLASS)2} = 0$ and $b_{1(CLASS)3} = 1$. Likewise, its value for *SEX* is solely explained by $SEX = 0$ such that $b_{1(SEX)1} = 1$ and $b_{1(SEX)2} = 0$. The continuous feature AGE with value 22 is explained by $AGE > y$ for $y =$ 14.0, 19.0 and by $AGE \leq x$ for $x =$ 22.0, 25.0, 28.0, 31.0, 36.0, 41.0, 50.0. The remainder of the rules for AGE does not explain the first observation. This thus results in the respective values $b_{1(AGE)n}$ of 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0 for $n =$ 1, ..., 18 respectively. A similar reasoning applies to the continuous features *NSIB*, $NPAT$ and *FARE*. The categorical feature *EMB* conforms to the processes of *CLASS* and *SEX*. In a similar fashion, the values for the second observation can be derived. The values of the parameter b_{isn} for the two observations displayed in Table 5 can be found in Table 6.

It is important to note that the explanations for the ordinal *CLASS* feature change for the extended unweighted Gower distance measure. Then, for the first observation, $b_{1(CLASS)1} = 0$, $b_{1(CLASS)2} = 1$, $b_{1(CLASS)3} = 0$, $b_{1(CLASS)4} = 1$, $b_{1(CLASS)5} = 1$ and $b_{1(CLASS)6} = 0$, conforming to the rules $CLASS \leq 1$, $CLASS > 1$, $CLASS \leq 2$, $CLASS > 2$, $CLASS \leq 3$, $CLASS > 3$.

Table 6: Explanations of the first two observations of the *titanic* dataset

Observation	Feature	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Observation 1	CLASS	0	0	1															
	SEX	1	0																
	AGE	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0
	NSIB	0	1	1	0														
	NPAT	1	0	1	0	1	0												
	FARE	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
	EMB	0	0	1															
Observation 2	CLASS	1	0	0															
	SEX	0	1																
	AGE	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	1	0
	NSIB	0	1	1	0														
	NPAT	1	0	1	0	1	0												
	FARE	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0
	EMB	1	0	0															

Note. The value of one (zero) represents that that observation is (not) explained by rule n of feature s. If there is no value, that rule does not exist.

6.3 InterP

The results of InterP, with a maximum of two joined explanations per cluster, for $\theta = 1$ are displayed in Table 7 and for $\theta = 4$, $\theta = 2$, $\theta = 0.5$ and $\theta = 0.25$ in Tables A7-A10. Looking at Table 7, the first cluster of the *wine* dataset, for example, conforms to the explanation $ALCOH > 12.77$ AND $FLAV > 2.14$, which yields a true positive rate of 100% and a false positive rate of 3%. Giving equal weight to the accuracy and distinctiveness, the true positive rate ranges from 6% to 100% and the false positive rate ranges from 0% to 28% across all clusters. A low true positive rate is the consequence of the trade-off between accuracy and distinctiveness since there exist rules that yield a higher true positive rate for that cluster (in an extreme case, for example, taking the rule being smaller than or equal to the ninth decile value) but then the false positive rate increases as least as much as this solution does not conform to the optimal one (within the time limit).

On an extreme end, $\theta = 4$ corresponds to the case that minimal weight is given to explanations within the cluster that satisfy the cluster's explanation (i.e., $\theta_1 = 0.5$) and maximal weight to explanations outside the cluster that satisfy the cluster's explanation (i.e., $\theta_2 = 2$). On the other side of the spectrum, $\theta = 0.25$ corresponds to the case that $\theta_1 = 2$ and $\theta_2 = 0.5$. This pattern is also visible in the results. As the value of θ increases (decreases), the value of the true positive rate for the clusters overall decreases (increases), but also at the gain (cost) of a lower (higher) false positive rate since more (less) weight is attached to them. The relatively small differences in the true positive rates and false positive rates compared to the results of Carrizosa et al. (2023) are likely the cause of differences in the set of rules, but it is not possible to be conclusive about this without insights into the rules they used. The sizes of the set of rules did however differ between this paper and their paper, which supports the argument.

Table 7: InterP clusters and cluster explanations for the various datasets with $\theta = 1$ and explanations of a maximum length of 2

Dataset	Cluster	Performance		Explanations
		TPR	FPR	
wine	1	1.00 (1.00)	0.03 (0.03)	$ALCOH > 12.77$ AND $FLAV > 2.14$
	2	0.86 (0.83)	0.02 (0.01)	$ALCOH \leq 12.77$ AND $FLAV > 0.84$
	3	0.94 (0.98)	0.01 (0.02)	$FLAV \leq 1.32$ AND $COLINT > 3.4$
glass	1	0.76 (0.76)	0.17 (0.17)	$RI > 1.51735$ AND $Mg > 3.39$
	2	0.54 (0.54)	0.12 (0.12)	$Mg > 2.81$ AND $Ca > 8.12$
	3	0.06 (0.06)	0.00 (0.00)	$Na > 14.03$ AND $Fe > 0.14$
	4	0.54 (0.23)	0.01 (0.00)	$Al > 1.36$ AND $Ca > 10.56$
	5	1.00 (0.67)	0.02 (0.01)	$K \leq 0.0$ AND $Ba \leq 0.0$
	6	0.79 (0.79)	0.00 (0.00)	$Na > 14.03$ AND $Ba > 0.0$
housing	1	0.76 (0.70)	0.05 (0.06)	$RM > 6.086$ AND $LSTAT \leq 11.38$
	2	0.82 (0.81)	0.15 (0.23)	$INDUS > 4.39$ AND $LSTAT > 9.53$
abalone	1	0.73 (0.71)	0.05 (0.18)	$WEIGHT \leq 0.13$ AND $SHEWEIG \leq 0.14$
	2	0.94 (0.76)	0.24 (0.23)	$WHWEIG > 0.45$ AND $SHEWEIG > 0.14$
contraceptive	1	0.28	0.28	$WAGE > 37.0$ AND $WEDUC \leq 3.0$
	2	0.08	0.07	$WEDUC > 3.0$ AND $CHILD > 3.0$
	3	0.32	0.13	$WAGE \leq 32.0$ AND $CHILD > 2.0$
titanic	1	0.81	0.25	$AGE > 14.0$
	2	0.66	0.13	$SEX = 1.0$ AND $NPAT \leq 2$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

6.4 CinterP

The results of CinterP, with a maximum of two joined explanations per cluster, for $\theta_1 = 0.5$ and $\theta_2 = 0.5$ and for the squared Euclidean distance (SED), unweighted Gower distance (UG) and extended unweighted Gower distance (EUG) are presented in Table 8. Since $\theta_1 = \theta_2$, accuracy and distinctiveness have equal importance. The *housing* dataset, for example, can be explained by $RAD \leq 8.0$ AND $TAX \leq 666.0$ for the first cluster and $TAX > 437.0$ AND $PTRATIO \leq 20.2$ for the second cluster, yielding a total distance between the observations of $1.5 \cdot 10^6$, twice a true positive rate of 100% and twice a false positive rate of 0%. The results for the various datasets of the combinations (0.5,1), (0.5,2), (1,0.5) and (2,0.5) for (θ_1,θ_2) are displayed in Tables A11-A14.

Table 8: CinterP clusters and cluster explanations for the various datasets with $\theta_1 = 0.5$ and $\theta_2 = 0.5$ and explanations of a maximum length of 2

Dataset	Distance	Cluster	Performance			Explanations
			Intra-homogeneity	TPR	FPR	
wine	SED	1	$1.2 \cdot 10^5$ $(5.0 \cdot 10^3)$	1.00 (1.00)	0.00 (0.00)	$ALCOH \leq 12.77$ AND $PHENF > 0.26$
		2		1.00 (1.00)	0.00 (0.00)	$PHENF \leq 0.26$
		3		1.00 (1.00)	0.00 (0.00)	$ALCOH > 12.77$ AND $PHENF > 0.26$
	UG	1	$1.8 \cdot 10^4$	1.00	0.00	$HUE > 1.24$
		2		1.00	0.00	$FLAV \leq 1.32$ AND $HUE \leq 1.24$
		3		1.00	0.00	$FLAV > 1.32$ AND $HUE \leq 1.24$
	EUG	1	$4.2 \cdot 10^4$	1.00	0.00	$PROL > 1265.0$
		2		1.00	0.00	$ODODW \leq 2.52$
		3		1.00	0.00	$PROL \leq 1265.0$ AND $ODODW > 2.52$
glass	SED	1	$4.5 \cdot 10^4$ $(7.8 \cdot 10^2)$	0.07 (0.77)	0.00 (0.03)	$Al \leq 0.87$ AND $K > 0.19$
		2		0.95 (1.00)	0.03 (0.00)	$RI > 1.51869$ AND $Mg > 2.81$
		3		0.89 (0.95)	0.11 (0.08)	$RI > 1.52211$ AND $Mg \leq 0.0$
		4		0.58 (1.00)	0.25 (0.01)	$RI \leq 1.52211$ AND $Ca > 10.56$
		5		0.78 (0.96)	0.14 (0.01)	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.88 (0.44)	0.04 (0.00)	$Na > 14.03$ AND $Si > 72.79$
	UG	1	$5.6 \cdot 10^3$	0.14	0.00	$Al \leq 0.87$ AND $K > 0.19$
		2		0.93	0.03	$RI > 1.51869$ AND $Mg > 2.81$
		3		0.89	0.11	$RI > 1.52211$ AND $Mg \leq 0.0$
		4		0.54	0.23	$RI \leq 1.52211$ AND $Ca > 10.56$
		5		0.74	0.13	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.85	0.04	$Na > 14.03$ AND $Si > 72.79$
	EUG	1	$3.3 \cdot 10^4$	0.14	0.00	$Al \leq 0.87$ AND $K > 0.19$
		2		0.93	0.03	$RI > 1.51869$ AND $Mg > 2.81$
		3		0.89	0.11	$RI > 1.52211$ AND $Mg \leq 0.0$
		4		0.54	0.23	$RI \leq 1.52211$ AND $Ca > 10.56$
		5		0.74	0.13	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.85	0.04	$Na > 14.03$ AND $Si > 72.79$
housing	SED	1	$1.5 \cdot 10^6$ $(6.0 \cdot 10^4)$	1.00 (1.00)	0.00 (0.04)	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00 (1.00)	0.00 (0.00)	$TAX > 437.0$ AND $PTRATIO \leq 20.2$
	UG	1	$1.6 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$TAX > 437.0$ AND $PTRATIO \leq 20.2$
	EUG	1	$4.0 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$TAX > 437.0$ AND $PTRATIO \leq 20.2$
abalone	SED	1	$3.1 \cdot 10^6$ $(2.2 \cdot 10^5)$	1.00 (0.82)	0.00 (0.16)	$SEX = 2.0$ AND $SHEWEIG \leq 0.12$
		2		1.00 (1.00)	0.17 (0.00)	$DIAM > 0.255$
	UG	1	$5.9 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		1.00	0.17	$DIAM > 0.255$
	EUG	1	$7.4 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		1.00	0.17	$DIAM > 0.255$
contraceptive	SED	1	$1.7 \cdot 10^7$	0.66	0.27	$WAGE \leq 37.0$ AND $CHILD > 2.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	UG	1	$1.3 \cdot 10^6$	0.66	0.27	$WAGE \leq 37.0$ AND $CHILD > 2.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	EUG	1	$1.4 \cdot 10^6$	0.66	0.27	$WAGE \leq 37.0$ AND $CHILD > 2.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
titanic	SED	1	$6.1 \cdot 10^6$	1.00	0.00	$NPAT \leq 2.0$
		2		1.00	0.00	$NPAT > 2.0$
	UG	1	$4.2 \cdot 10^5$	0.97	0.00	$FARE \leq 79.2$
		2		1.00	0.46	$FARE > 79.2$
	EUG	1	$8.7 \cdot 10^5$	1.00	0.00	All in
		2		0.00	0.00	-

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

A notable difference between the results in this paper and the results of Carrizosa et al. (2023) is the difference in intra-homogeneity. These differences are likely due to the difference in the normalization technique applied between the two papers. Carrizosa et al. (2023) do namely not state which normalization technique they use. For example, using min-max normalization rather than Z-score normalization, the intra-homogeneity of the *housing* dataset for $\theta_1 = 0.5$ and $\theta_2 = 0.5$ drops to $8.1 \cdot 10^4$, considering SED, while keeping similar performance regarding the true positive rates and false positive rates. For the *wine* dataset, the intra-homogeneity then drops to $7.8 \cdot 10^2$, while also having similar performance with regard to the TPRs and FPRs. For the *glass* dataset, the intra-homogeneity then drops to $1.3 \cdot 10^3$ and the TPRs and FPRs become similar to that of UG and EUG. Since the cluster allocations and the precise set of rules of Carrizosa et al. (2023) are unknown, it cannot be said of part of the differences in the performance measures is also due to different cluster allocations or differences in the set of rules. The TPRs and FPRs are, however, similar which indicates that the cluster allocations cannot be totally different.

Another difference between this paper and the paper of Carrizosa et al. (2023) concerns the definitions of α_i and β_{ki} in the MILP of CinterP. Although the binary restrictions used in this paper are technically the correct definitions, Carrizosa et al. (2023) use the interval domain [0,1], as this speeds up the process of finding the optimal solution, especially considering there is a time limit, since the constraints ensure these variables will eventually either take the value zero or one. The results of Table 8 only change for the *wine* dataset if the time limit is reduced from ten to five minutes, which mitigates the possibility of the differences in the definitions being the main cause of the differences in results.

Despite the differences in the set of rules, the definitions of α_i and β_{ki} and (most probably) the normalization technique between the two papers, the implementation can be considered to be successful since the true positive rate and false positive rate of the clusters are similar and interpretability of the clusters is ensured. The differences in intra-homogeneity are most likely due to different choices in normalization techniques, and thus the consequence of an assumption that is made rather than a difference in the method used.

6.5 Distance measures

Since distances between observations are measured differently for SED, UG and EUG, observations can be allocated differently respective to the distance measure under consideration. For example, although the selection of rules for the *glass* dataset in Table 8 is exactly the same for SED, UG and EUG, the performance measures alter, indicating different cluster allocations (as otherwise the true positive rate and false positive rate have to be the same across SED, UG and EUG, despite having a different intra-homogeneity). Since distances between observations are measured differently for the different distance measures, the results for the various datasets across SED, UG and EUG can only be compared based on their TPR and FPR.

For the datasets containing continuous features only, meaning the *wine* and *glass* datasets, the performance across the distance measures and combinations of θ_1 and θ_2 are mostly similar. In some cases, however, UG and EUG perform better than SED in allocating the observations to clusters and finding rules for the corresponding clusters, like for the *wine* dataset with $\theta_1 = 1$ and $\theta_2 = 0.5$ in Table A13. Performing better (worse) means, for all clusters, the TPR is at least as high while the FPR did not increase or the FPR is at most as high while the TPR did not decrease. For the different distance measures concerning the *glass* dataset with $\theta_1 = 0.5$ and $\theta_2 = 0.5$, see Table 8, some clusters yield a better TPR and FPR but other clusters have worse TPR and FPR at the same time. However, for any combination of θ_1 and θ_2, SED never performs better than UG or EUG. The only difference between SED on the one hand and UG and EUG on the other hand, having continuous features only, is the scaling of these continuous features. SED namely uses Z-score normalization whereas UG and EUG use range-normalization with the Manhattan distance.

For the *housing* and *abalone* datasets that contain continuous features and a single categorical feature, no differences regarding the true and false positive rates are found for any of the θ_1 and θ_2 combinations. Moreover, six out of ten times, SED, UG and EUG provided the exact same cluster explanations, whereas the other four times the explanations of SED were longer than those of UG and EUG, namely $SEX = 2.0$ AND $SHEWEIG \leq 0.12$ for the first cluster of the *abalone* dataset using SED compared with $SEX = 2.0$ for the first cluster using UG and EUG, indicating a benefit of UG and EUG over SED regarding interpretability. In this case, the difference in measuring the distance between categorical features across SED on the one hand and UG and EUG on the other hand might be flattered since there are many more continuous features than categorical features in both datasets. If, for instance, only the first four features of the *housing* dataset would be considered, meaning three continuous and one categorical feature instead of twelve continuous and one categorical feature, EUG performs better than SED and UG. Not only does EUG obtain better TPRs and FRPs (100% and 0% for both clusters compared to 76% and 0% for the first cluster and 100% and 5% for the second cluster using SED and UG) but it also provides shorter cluster explanations. This result is in line with the statement of D'Orazio (2021) regarding the unbalanced contribution becoming more apparent when the relative number of categorical features increases. The relatively more categorical features a dataset has, the more the benefit of EUG should become evident.

Regarding the *contraceptive* dataset having two continuous, three dummy and four ordinal features, EUG never performs worse than the SED or UG. Having equal true and false positive weights or more weight towards the true positive rate, all three distance measures yield the same TPR, FPR and cluster explanations. When more weight is put towards the false positive rate, EUG either performs better or provides a shorter explanation while having similar TPRs and FPRs. For $\theta_1 = 0.5$ and $\theta_2 = 1$, SED performs worse than UG and EUG. The *titanic* dataset contains four continuous features, one dummy feature, one categorical feature and one ordinal feature. Rather surprisingly, SED performs the best for any combination of θ_1 and θ_2 since it either separates the observations into clusters whereas EUG cannot or has higher true positive rates or lower false positive rates compared to UG. However, if a different seed for the MIP start is chosen, SED, UG and EUG all yield true positive rates of 97% and 100%, false positive rates of 0% and respectively 46% and the same cluster explanations. This result thus shows the importance of a good initial solution. Hereby is to say that the k-means algorithm is based on the Euclidean distance metric as well. Additionally, since the Euclidean distance metric is unable to handle missing values, observations containing missing values were deleted from the *titanic* dataset. This, however, mitigates the ability of the Gower distance to account for missing values. If these observations were not deleted, and using the original seed, EUG is able to divide the titanic dataset into two clusters. In general, values could also be missing for a reason, so simply deleting observations containing missing values could result in biased estimates and therefore reduces the statistical power of the analysis, which also implies a benefit of the (extended) unweighted Gower distance over the squared Euclidean distance.

In conclusion, with advanced initialization, the extended unweighted Gower distance provides the most promising results for allocating observations to clusters and finding cluster explanations. Either it has a similar performance to the squared Euclidean distance and unweighted Gower distance or it outperforms (one of the) two or it has a similar performance but shorter explanations. Moreover, the extended unweighted Gower distance can handle missing data.

7 Conclusion

Although the importance of interpretable clustering is well-understood (Doshi-Velez & Kim, 2017), most clustering algorithms focus more on performance than interpretability (Bertsimas et al., 2021). Therefore, Carrizosa et al. (2023) introduced a new approach of explaining clusters in which the explanations are based on the observation's characteristics, combined with the AND operator. Two different MILPs are considered, one for finding explanations when clusters are already given, referred to as InterP, and one for simultaneously finding clusters and their explanations, referred to as CinterP. The performance measures considered are the sum of distances between observations within the same cluster (i.e., intra-homogeneity), the rate of observations within the cluster satisfying the cluster's explanation (i.e., accuracy) and the rate of observations outside the cluster satisfying the cluster's explanation (i.e., distinctiveness). Since the approach of Carrizosa et al. (2023) considers the squared Euclidean distance, CinterP is limited to a distance measure suitable for continuous features only (Lee, 2022), whereas most datasets consist of multiple feature types. This paper, therefore, considers two forms of the more widely applicable Gower distance (Gower, 1971), namely the unweighted Gower distance and the extended unweighted Gower distance. The extended unweighted Gower distance also accounts for ordinal features (Podani, 1999) and tackles the unequal contribution of features the unweighted Gower distance faces (D'Orazio, 2021).

Based on six datasets that differ in their feature types, the squared Euclidean distance, the unweighted Gower distance and the extended unweighted Gower distance are compared. This comparison is based on the accuracy and distinctiveness of the clusters as well as the cluster explanations. If yielding unequal results, the use of the extended unweighted Gower distance results in better cluster allocations or shorter cluster explanations compared to the squared Euclidean distance and the unweighted Gower distance, implying the extended unweighted Gower distance is the preferred distance measure for the MILP formulation of CinterP.

To end, this research can be improved in several ways. Firstly, it would be interesting to look into the initial solution for the MIP start. For example, compared to the k-means algorithm, the Gaussian Mixture model does not need circular-shaped data to work well and DBSCAN can handle datasets with varying densities and cluster sizes or identify clusters with arbitrary shapes (McGregor, 2020). Secondly, since the unweighted Gower distance assigns equal weight to all features, clusters can be dominated by one type of feature type (Hendrickson, 2014). Assigning different weights may reduce this problem. Thirdly, the extended unweighted Gower distance can be further extended with other types of features, like circular and proportion features (Pavoine, Vallet, Dufour, Gachet & Daniel, 2009) or the partial distances can be measured differently (Šulc, Matějka, Procházka & Řezanková, 2017). Lastly, restricting features to explain a single cluster only or modeling fairness constraints are other tracks of research worth considering.

References

Aggarwal, C. C., Hinneburg, A. & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Database theory—icdt 2001: 8th international conference london, uk, january 4–6, 2001 proceedings 8* (pp. 420–434).

Arat, M. (2019). *A complete guide to k-nearest-neighbors with applications in python.* Retrieved from `https://mmuratarat.github.io/2019-07-12/k-nn-from-scratch#:~:text=K%2DNN%20algorithm%20is%20an,samples%20in%20the%20training%20dataset`

Bertsimas, D., Orfanoudaki, A. & Wiberg, H. (2021). Interpretable clustering: an optimization approach. *Machine Learning*, *110*, 89–138.

Boehmke, B. & Greenwell, B. M. (2019). *Hands-on machine learning with r.* CRC press.

Carrizosa, E., Kurishchenko, K., Marín, A. & Romero Morales, D. (2023). On clustering and interpreting with rules by means of mathematical optimization. *Computers Operations Research*, *154*, 106180. doi: https://doi.org/10.1016/j.cor.2023.106180

Chen, J. (2018). *Interpretable clustering methods* (Unpublished doctoral dissertation). Northeastern University.

Cplex, I. I. (2009). V12. 1: User's manual for cplex. *International Business Machines Corporation*, *46*(53), 157.

Cui, X., Potok, T. E. & Palathingal, P. (2005). Document clustering using particle swarm optimization. In *Proceedings 2005 ieee swarm intelligence symposium, 2005. sis 2005.* (pp. 185–191).

D'Orazio, M. (2021). *Distances with mixed type variables some modified gower's coefficients.*

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dua, D. & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Foss, A. H. & Markatou, M. (2018). kamila: clustering mixed-type data in r and hadoop. *Journal of Statistical Software*, *83*, 1–44.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Hajnal, I. & Loosveldt, G. (1998). An evaluation of some clustering methods for mixed mode variable data sets. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *58*(1), 16–30.

Harrison, E. & Pius, R. (2020). *R for health data science.* CRC Press.

Hendrickson, J. (2014). *Methods for clustering mixed data* (Unpublished doctoral dissertation).

University of South Carolina.

IBM. (2022). *Starting from a solution: Mip starts.* Retrieved from https://www.ibm.com/docs/en/icos/22.1.0?topic=mip-starting-from-solution-starts

Jang, W. & Hendry, M. (2007). Cluster analysis of massive datasets in astronomy. *Statistics and Computing*, *17*, 253–262.

Kenny, E. M., Delaney, E. D., Greene, D. & Keane, M. T. (2021). Post-hoc explanation options for xai in deep learning: The insight centre for data analytics perspective. In *Pattern recognition. icpr international workshops and challenges: Virtual event, january 10–15, 2021, proceedings, part iii* (pp. 20–34).

Kolbe-Alexander, T. L., Conradie, J. & Lambert, E. V. (2013). Clustering of risk factors for non-communicable disease and healthcare expenditure in employees with private health insurance presenting for health risk appraisal: a cross-sectional study. *BMC public health*, *13*(1), 1–10.

Lasaosa, J. M. (2021). *Clustering on numerical and categorical features.* Retrieved from https://towardsdatascience.com/clustering-on-numerical-and-categorical-features-6e0ebcf1cbad

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. & Detyniecki, M. (2019). The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*.

Lawless, C., Kalagnanam, J., Nguyen, L. M., Phan, D. & Reddy, C. (2022). Interpretable clustering via multi-polytope machines. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 7309–7316).

Lee, J. (2022). *When do we not use euclidean distance?* Retrieved from https://jedleee.medium.com/when-do-we-not-use-euclidean-distance-dd9009ddda43

Løvik, I. K., Siglen, E. & Bjorvatn, C. (2022). Adaptive translation of the genetic counselling outcome scale–.

McGregor, M. (2020). *8 clustering algorithms in machine learning that all data scientists should know.* Retrieved from https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

Mihova, V. & Pavlov, V. (2018). A customer segmentation approach in commercial banks. In *Aip conference proceedings* (Vol. 2025, p. 030003).

Mimmack, G. M., Mason, S. J. & Galpin, J. S. (2001). Choice of distance matrices in cluster analysis: Defining regions. *Journal of climate*, *14*(12), 2790–2797.

Mohamad, I. B. & Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303.

Nagwani, N. K. & Sharaff, A. (2017). Sms spam filtering and thread identification using bi-level text classification and clustering techniques. *Journal of Information Science*, *43*(1), 75–87.

openletter.mousetail.nl. (2019). *How does a warm start work in lp/mip?* Retrieved from `https://or.stackexchange.com/questions/1278/how-does-a-warm-start-work-in-lp-mip`

Palmer, E. (2019). *Ordinal, nominal, ... who cares?* Retrieved from `https://towardsdatascience.com/ordinal-nominal-who-cares-82c867d7b774`

Pavoine, S., Vallet, J., Dufour, A.-B., Gachet, S. & Daniel, H. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, *118*(3), 391–402.

Petchey, O. L. & Gaston, K. J. (2007). Dendrograms and measuring functional diversity. *Oikos*, *116*(8), 1422–1426.

Plant, C. & Böhm, C. (2011). Inconco: interpretable clustering of numerical and categorical objects. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1127–1135).

Podani, J. (1999). Extending gower's general coefficient of similarity to ordinal characters. *Taxon*, *48*(2), 331–340.

Prabakaran, S. & Mitra, S. (2018). Survey of analysis of crime detection techniques using data mining and machine learning. In *Journal of physics: Conference series* (Vol. 1000, p. 012046).

Ramirez-Cano, D., Colton, S. & Baumgarten, R. (2010). Player classification using a meta-clustering approach. In *Proceedings of the 3rd annual international conference computer games, multimedia & allied technology* (pp. 297–304).

Saxena, A. & Singh, M. (2016). Using categorical attributes for clustering. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *2*(2), 324–329.

Šulc, Z., Matějka, M., Procházka, J. & Řezanková, H. (2017). Evaluation of the gower coefficient modifications in hierarchical clustering. *Advances in Methodology and Statistics*, *14*(1), 37–48.

Toragay, O. (2019). *What are the benefits of linearization?* Retrieved from `https://or.stackexchange.com/questions/2892/what-are-the-benefits-of-linearization`

Tran, L., Fan, L. & Shahabi, C. (2021). Clustering mixed-type data with correlation-preserving embedding. In *Database systems for advanced applications: 26th international conference, dasfaa 2021, taipei, taiwan, april 11–14, 2021, proceedings, part ii 26* (pp. 342–358).

van Fraassen, F., Hensen, C., de Wind, M. & van Exel, F. (2023). *Home sweet home: Predicting housing prices with machine learning.*

Zhang, C., Gupta, A., Kauten, C., Deokar, A. V. & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research*, *279*(3), 1036-1052. doi: https://doi.org/10.1016/j.ejor.2019.06.022

8 Appendix

Table A1: Description of the wine dataset with 178 observations and 3 classes

Feature	Description	Feature type
ALCOH	Alcohol	Continuous
MALAC	Malic acid	Continuous
ASH	Ash	Continuous
ALCASH	Alcalinity of ash	Continuous
Mg	Magnesium	Continuous
PHEN	Total phenols	Continuous
FLAV	Flavanoids	Continuous
PHENF	Nonflavanoid phenols	Continuous
PROAN	Proanthocyanins	Continuous
COLINT	Color intensity	Continuous
HUE	Hue	Continuous
ODODW	OD280/OD315 of diluted wines	Continuous
PROL	Proline	Continuous
	Type of wine (class 1, 2 or 3)	

Table A2: Description of the glass dataset with 214 observations and 6 classes

Feature	Description	Feature type
Ri	Refractive index	Continuous
Na	Sodium	Continuous
Mg	Magnesium	Continuous
Al	Aluminium	Continuous
Si	Silicon	Continuous
K	Potassium	Continuous
Ca	Calcium	Continuous
Ba	Barium	Continuous
Fe	Iron	Continuous
Class	Type of glass (class 1, 2, 3, 4, 5 or 6)	

Table A3: Description of the housing dataset with 392 observations and 2 classes

Feature	Description	Feature type
CRIM	per capita crime rate by town	Continuous
ZONE	proportion of residential land zoned for lots over 25,000 sq.ft.	Continuous
INDUS	proportion of non-retail business acres per town	Continuous
CHAS	Indicates whether the Charles River is nearby (1 if tract bounds river, 0 otherwise)	Dummy
NOXID	nitric oxides concentration (parts per 10 million)	Continuous
ROOMS	average number of rooms per dwelling	Continuous
AGE	proportion of owner-occupied units built prior to 1940	Continuous
DIST	weighted distances to five Boston employment centres	Continuous
RAD	index of accessibility to radial highways	Continuous
TAX	full-value property-tax rate per \$10,000	Continuous
PTRATIO	pupil-teacher ratio by town	Continuous
BPROP	$1000(Bk\&0.63)^2$ where Bk is the proportion of blacks by town	Continuous
LSTAT	% lower status of the population	Continuous
Class	Higher (class 0) or lower (class 1) than the median value of the owner-occupied homes in $1000's	

Note. Features are displayed to their most specific type, i.e. a dummy is more specific than categorical.

Table A4: Description of the abalone dataset with 4177 observations and 2 classes

Feature	Description	Feature type
SEX	Sex (0 if male, 1 if female, 2 if infant)	Categorical
LENGHT	Length	Continuous
DIAM	Diameter	Continuous
HEIGHT	Height	Continuous
WHWEIG	Whole weight	Continuous
SHUWEIG	Shucked weight	Continuous
VIWEIG	Viscera weight	Continuous
SHEWEIG	Shell weight	Continuous
Class	Higher (class 2) or lower (class 1) than the median value of the number of the rings	

Table A5: Description of the contraceptive dataset with 1473 observations and 3 classes

Feature	Description	Feature type
WAGE	Wife's age	Continuous
WEDUC	Wife's education (1 if low, 2 if medium-low, 3 if medium-high, 4 if high)	Ordinal
HEDUC	Husband's education (1 if low, 2 if medium-low, 3 if medium-high, 4 if high)	Ordinal
CHILD	Number of children ever born	Continuous
RELIG	Wife's religion (1 if Islam, 0 otherwise	Dummy
WORK	Indicates whether the wife is working now (1 if working, 0 otherwise)	Dummy
HOCC	Husband's occupation (1, 2, 3 or 4)	Ordinal
STLIV	Standard-of-living (1 if low, 2 if medium-low, 3 if medium-high, 4 if high)	Ordinal
MEDEXP	Media exposure (1 if not good, 0 if good)	Dummy
Class	No (class 1), long-term (class 2) or short-term (class 3) contraceptive use	

Note. Features are displayed to their most specific type, i.e. a ordinal is more specific than categorical.

Table A6: Description of the titanic dataset with 712 observations and 2 classes

Feature	Description	Feature type
CLASS	Passenger class (1 if first, 2 if second, 3 if third)	Ordinal
SEX	Sex (0 if male, 1 if female)	Dummy
AGE	Age	Continuous
NSIB	Number of siblings & spouses of the passenger aboard at the Titanic	Continuous
NPAT	Number of parents & children of the passenger aboard at the Titanic	Continuous
FARE	Ticket price paid in ponds	Continuous
EMB	Port of Embarkation (0 if Cherbourg, 1 if Queenstown, 2 if Southampton))	Categorical
Class	Passenger did not survive (class 0) or did survive (class 1)	

Table A7: InterP clusters and cluster explanations for the various datasets with $\theta = 4$ and explanations of a maximum length of 2

Dataset	Cluster	Performance		Explanations
		TPR	FPR	
wine	1	0.97 (0.86)	0.02 (0.01)	$FLAV > 2.14$ AND $PROL > 740.0$
	2	0.85 (0.77)	0.01 (0.00)	$ALCOH \leq 12.77$ AND $COLINT \leq 4.7$
	3	0.88 (0.90)	0.00 (0.00)	$FLAV \leq 1.32$ AND $COLINT > 4.1$
glass	1	0.10 (0.14)	0.00	$Ri > 1.52211$ AND $Mg > 3.54$
	2	0.13 (0.14)	0.00 (0.01)	$Mg > 3.76$ AND $Ca \leq 8.48$
	3	0.06 (0.06)	0.00 (0.00)	$Na > 14.03$ AND $Fe > 0.22$
	4	0.23 (0.23)	0.00 (0.00)	$Ri \leq 1.5167$ AND $Si \leq 71.77$
	5	0.22 (0.22)	0.00 (0.00)	$K \leq 0.0$ AND $Ca \leq 7.97$
	6	0.79 (0.79)	0.00 (0.00)	$Na > 14.03$ AND $Ba > 0.0$
housing	1	0.68 (0.64)	0.02 (0.01)	$ROOMS > 6.086$ AND $LSTAT \leq 9.53$
	2	0.66 (0.45)	0.04 (0.05)	$TAX > 289.0$ AND $LSTAT > 13.33$
abalone	1	0.49 (0.50)	0.02 (0.04)	$WHWEIG \leq 0.2715$ AND $VIWEIG \leq 0.06$
	2	0.70 (0.42)	0.05 (0.05)	$LENGTH > 0.525$ AND $SHEWEIG > 0.235$
contraceptive	1	0.10	0.01	$WAGE > 32.0$ AND $CHILD \leq 1.0$
	2	0.00	0.00	$WAGE > 45.0$ AND $WEDUC > 4.0$
	3	0.00	0.00	$WAGE > 45.0$ AND $WEDUC > 4$
titanic	1	0.54	0.12	$SEX = 0.0$ AND $FARE \leq 15.7417$
	2	0.28	0.01	$CLASS = 1.0$ AND $SEX = 1.0$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A8: InterP clusters and cluster explanations for the various datasets with $\theta = 2$ and explanations of a maximum length of 2

Dataset	Cluster	Performance		Explanations
		TPR	FPR	
wine	1	0.97 (1.00)	0.02 (0.03)	$FLAV > 2.14$ AND $PROL > 740.0$
	2	0.85 (0.83)	0.01 (0.01)	$ALCOH \leq 12.77$ AND $COLINT \leq 4.7$
	3	0.94 (0.90)	0.01 (0.00)	$FLAV \leq 1.32$ AND $COLINT > 3.4$
glass	1	0.43 (0.43)	0.07 (0.07)	$Mg > 3.39$ AND $Ca > 8.6$
	2	0.33 (0.33)	0.04 (0.04)	$Mg > 3.48$ AND $Ca \leq 8.12$
	3	0.06 (0.06)	0.00 (0.00)	$Na > 14.03$ AND $Fe > 0.14$
	4	0.23 (0.23)	0.00 (0.00)	$Ri \leq 1.5167$ AND $Si \leq 71.77$
	5	0.22 (0.22)	0.00 (0.00)	$K \leq 0.0$ AND $Ca \leq 7.97$
	6	0.79 (0.79)	0.00 (0.00)	$Na > 14.03$ AND $Ba > 0.0$
housing	1	0.76 (0.70)	0.05 (0.04)	$ROOMS > 6.086$ AND $LSTAT \leq 11.38$
	2	0.66 (0.70)	0.04 (0.14)	$TAX > 289.0$ AND $LSTAT > 13.33$
abalone	1	0.72 (0.50)	0.05 (0.04)	$WEIGHT \leq 0.13$ AND $SHEWEIG \leq 0.12$
	2	0.85 (0.65)	0.15 (0.14)	$LENGTH > 0.485$ AND $WEIGHT > 0.115$
contraceptive	1	0.17	0.02	$WAGE > 27.0$ AND $CHILD \leq 1.0$
	2	0.00	0.00	$WEDUC > 4.0$
	3	0.01	0.00	$WAGE \leq 22.0$ AND $RELIG = 0.0$
titanic	1	0.81	0.25	$SEX = 0.0$ AND $AGE > 14.0$
	2	0.66	0.13	$SEX = 1.0$ AND $NPAT \leq 2$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A9: InterP clusters and cluster explanations for the various datasets with $\theta = 0.5$ and explanations of a maximum length of 2

Dataset	Cluster	Performance		Explanations
		TPR	FPR	
wine	1	1.00 (1.00)	0.03 (0.03)	$ALCOH > 12.77$ AND $FLAV > 2.14$
	2	0.86 (0.89)	0.02 (0.07)	$ALCOH \leq 12.77$ AND $FLAV > 0.84$
	3	1.00 (1.00)	0.03 (0.03)	$FLAV \leq 1.75$ AND $COLINT > 3.4$
glass	1	0.84	0.23	$Ri > 1.51735$ AND $Mg > 2.81$
	2	0.62	0.20	$Mg > 2.81$ AND $Ca \leq 8.48$
	3	0.12	0.01	$Na > 13.44$ AND $Fe > 0.22$
	4	0.92	0.05	$Na \leq 13.44$ AND $Mg \leq 2.81$
	5	1.0	0.02	$K \leq 0.0$ AND $Ba \leq 0.0$
	6	0.90	0.02	$Na > 13.3$ AND Al 1.76
housing	1	0.86 (0.78)	0.15 (0.19)	$ROOMS > 5.95$ AND $LSTAT \leq 13.33$
	2	0.99 (0.99)	0.42 (0.44)	$LSTAT \leq 7.74$
abalone	1	0.93 (0.88)	0.24 (0.41)	$LENGTH \leq 0.525$ AND $WEIGHT \leq 0.16$
	2	0.94 (0.86)	0.24 (0.34)	$WHWEIG >$0.45 AND $SHEWEIG > 0.14$
contraceptive	1	0.72	0.52	$WEDUC \leq 3$ AND $WEDUC > 3$
	2	0.43	0.14	$CHILD > 2.0$
	3	0.68	0.40	$WAGE \leq 41.0$ AND $CHILD > 1.0$
titanic	1	0.85	0.32	$SEX = 0.0$
	2	0.68	0.15	$SEX = 1.0$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A10: InterP clusters and cluster explanations for the various datasets with $\theta = 0.25$ and explanations of a maximum length of 2

Dataset	Cluster	Performance		Explanations
		TPR	FPR	
wine	1	1.00 (1.00)	0.03 (0.03)	$ALCOH > 12.77$ AND $FLAV > 2.14$
	2	0.94 (0.94)	0.19 (0.17)	$COLINT \leq 4.7$ AND $PROL \leq 1050.0$
	3	1.00 (1.00)	0.03 (0.03)	$FLAV \leq 1.75$ AND $COLINT > 3.4$
glass	1	0.94	0.37	$Al \leq 1.49$ AND $Ca \leq 10.56$
	2	0.95	0.68	$Na \leq 14.03$ AND $Ba \leq 0.64$
	3	0.35	0.05	$Ri \leq 1.51735$ AND $Al \leq 1.36$
	4	0.92	0.05	$Na \leq 13.44$ AND $Mg \leq 2.81$
	5	1.0	0.02	$K \leq 0.0$ AND $Ba \leq 0.0$
	6	0.90	0.02	$Ba > 0.0$ AND $Fe \leq 0.14$
housing	1	0.94 (0.98)	0.36 (0.80)	$PTRATIO \leq 20.9$ AND $LSTAT \leq 15.69$
	2	0.99 (0.99)	0.42 (0.44)	$LSTAT > 7.74$
abalone	1	0.93 (1.00)	0.24 (0.74)	$LENGTH \leq 0.525$ AND $WEIGHT \leq 0.16$
	2	0.98 (0.97)	0.47 (0.63)	$DIAM > 0.295$ AND $SHEWEIG > 0.0925$
contraceptive	1	1.00	1.00	$WEDUC \leq 4$
	2	0.74	0.42	$WEDUC > 2$ AND $CHILD > 1.0$
	3	0.93	0.76	$WAGE \leq 37.0$
titanic	1	0.96	0.81	$FARE \leq 79.2$
	2	0.91	0.72	$FARE > 7.8958$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A11: CinterP clusters and cluster explanations for the various datasets with $\theta_1 = 0.5$ and $\theta_2 = 1$ and explanations of a maximum length of 2

Dataset	Distance	Cluster	Performance			Explanations
			Intra-homogeneity	TPR	FPR	
wine	SED	1	$1.5 \cdot 10^5$ $(5.2 \cdot 10^3)$	0.63 (1.00)	0.00 (0.00)	$PROL > 1265.0$
		2		1.00 (1.00)	0.05 (0.00)	$PROL \leq 740.0$
		3		0.22 (1.00)	0.00 (0.00)	$MALAC > 2.68$ AND $PROL > 740.0$
	UG	1	$1.7 \cdot 10^4$	1.00	0.00	$HUE > 1.24$
		2		1.00	0.00	$PHEN \leq 2.53$ AND $HUE \leq 1.24$
		3		1.00	0.00	$PHEN > 2.53$ AND $HUE \leq 1.24$
	EUG	1	$4.5 \cdot 10^4$	1.00	0.00	$PROL > 1265.0$
		2		1.00	0.00	$COLINT \leq 4.1$ AND $HUE \leq 1.24$
		3		1.00	0.00	$COLINT > 4.1$ AND $PROL \leq 1265.0$
glass	SED	1	$5.2 \cdot 10^4$ $(9.2 \cdot 10^2)$	0.14 (0.53)	0.00 (0.02)	$Al > 1.23$ AND $K > 0.19$
		2		0.93 (1.00)	0.03 (0.00)	$Ri \leq 1.51869$ AND $Mg > 2.81$
		3		0.88 (1.00)	0.11 (0.04)	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.23 (1.00)	0.00 (0.01)	$Na > 14.03$ AND $Ca > 10.56$
		5		0.74 (0.91)	0.13 (0.00)	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.81 (0.44)	0.00 (0.00)	$K \leq 0.08$ AND $Ba > 0.0$
	UG	1	$5.6 \cdot 10^3$	0.14	0.00	$Al > 1.23$ AND $K > 0.19$
		2		0.93	0.03	$Ri \leq 1.51869$ AND $Mg > 2.81$
		3		0.88	0.11	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.23	0.00	$Na > 14.03$ AND $Ca > 10.56$
		5		0.74	0.13	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.81	0.00	$K \leq 0.08$ AND $Ba > 0.0$
	EUG	1	$3.3 \cdot 10^4$	0.14	0.00	$Al > 1.23$ AND $K > 0.19$
		2		0.93	0.03	$Ri \leq 1.51869$ AND $Mg > 2.81$
		3		0.88	0.11	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.23	0.00	$Na > 14.03$ AND $Ca > 10.56$
		5		0.74	0.13	$Si \leq 72.12$ AND $Ca > 8.6$
		6		0.81	0.00	$K \leq 0.08$ AND $Ba > 0.0$
housing	SED	1	$1.5 \cdot 10^6$ $(6.0 \cdot 10^4)$	1.00 (0.91)	0.00 (0.00)	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00 (1.00)	0.00 (0.00)	$CRIM > 0.09849$ AND $TAX > 666.0$
	UG	1	$1.6 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
	EUG	1	$4.0 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
abalone	SED	1	$3.1 \cdot 10^6$ $(2.2 \cdot 10^5)$	1.00 (0.82)	0.00 (0.16)	$SEX = 2.0$ AND $SHEWEIG \leq 0.12$
		2		0.82 (1.00)	0.00 (0.00)	$LENGTH > 0.425$
	UG	1	$5.9 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		0.82	0.00	$LENGTH > 0.425$
	EUG	1	$7.4 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		0.82	0.00	$LENGTH > 0.425$
contraceptive	SED	1	$1.7 \cdot 10^7$	0.27	0.04	$WAGE \leq 37.0$ AND $CHILD > 4.0$
		2		0.86	0.00	$WAGE \leq 27.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	UG	1	$1.3 \cdot 10^6$	0.45	0.10	$WAGE \leq 37.0$ AND $CHILD > 3.0$
		2		0.86	0.00	$WAGE \leq 27.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	EUG	1	$1.4 \cdot 10^6$	0.45	0.10	$WAGE \leq 37.0$ AND $CHILD > 3.0$
		2		0.86	0.00	$WAGE \leq 27.0$
		3		1.00	0.11	$WAGE > 37.0$
titanic	SED	1	$6.1 \cdot 10^6$	1.00	0.00	$NPAT \leq 2.0$
		2		1.00	0.00	$NPAT > 2.0$
	UG	1	$4.2 \cdot 10^5$	0.97	0.00	$FARE \leq 79.2$
		2		0.58	0.15	$NPAT > 0.0$ AND $FARE > 79.2$
	EUG	1	$8.7 \cdot 10^5$	1.00	0.00	$CLASS \leq 3.0$
		2		-	-	$CLASS > 3.0$ AND $FARE \leq 12.875$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A12: CinterP clusters and cluster explanations for the various datasets with $\theta_1 = 0.5$ and $\theta_2 = 2$ and explanations of a maximum length of 2

Dataset	Distance	Cluster	Performance			Explanations
			Intra-homogeneity	TPR	FPR	
wine	SED	1	$1.8 \cdot 10^5$ $(6.2 \cdot 10^3)$	1.00 (1.00)	0.00 (0.00)	$Mg > 95.0$
		2		1.00 (1.00)	0.00 (0.00)	$ALCASH \leq 16.8$ AND $Mg \leq 95.0$
		3		1.00 (1.00)	0.00 (0.00)	$ALCASH > 16.8$ AND $Mg \leq 95.0$
	UG	1	$1.7 \cdot 10^4$	1.00	0.00	$HUE > 1.24$
		2		1.00	0.00	$PHEN \leq 2.53$ AND $HUE \leq 1.24$
		3		1.00	0.00	$PHEN > 2.53$ AND $HUE \leq 1.24$
	EUG	1	$5.0 \cdot 10^4$	1.00	0.00	$FLAV > 0.84$ AND $PROL > 1265.0$
		2		1.00	0.00	$ALCASH > 18.6$ AND $PROL \leq 1265.0$
		3		1.00	0.00	$ALCASH \leq 18.6$ AND $PROL \leq 1265.0$
glass	SED	1	$5.2 \cdot 10^4$ $(8.6 \cdot 10^2)$	0.14 (0.24)	0.00 (0.00)	$Al > 1.23$ AND $K > 0.19$
		2		0.88 (1.00)	0.03 (0.00)	$Mg > 2.81$ AND $Ca \leq 8.78$
		3		0.67 (1.00)	0.00 (0.04)	$Ri > 1.52211$ AND $Na \leq 12.68$
		4		0.15 (0.90)	0.00 (0.00)	$Ri \leq 1.51869$ AND $Ca > 9.57$
		5		0.26 (0.91)	0.00 (0.00)	$Mg > 3.76$ AND $Ca > 8.6$
		6		0.81 (0.40)	0.00 (0.00)	$K \leq 0.08$ AND $Ba > 0.0$
	UG	1	$5.6 \cdot 10^3$	0.14	0.00	$Al > 1.23$ AND $K > 0.19$
		2		0.88	0.03	$Mg > 2.81$ AND $Ca \leq 8.78$
		3		0.67	0.00	$Ri > 1.52211$ AND $Na \leq 12.68$
		4		0.15	0.00	$Ri \leq 1.51869$ AND $Ca > 9.57$
		5		0.26	0.00	$Mg > 3.76$ AND $Ca > 8.6$
		6		0.81	0.00	$K \leq 0.08$ AND $Ba > 0.0$
	EUG	1	$3.3 \cdot 10^4$	0.14	0.00	$Al > 1.23$ AND $K > 0.19$
		2		0.88	0.03	$Mg > 2.81$ AND $Ca \leq 8.78$
		3		0.67	0.00	$Ri > 1.52211$ AND $Na \leq 12.68$
		4		0.08	0.00	$Ri \leq 1.5167$ AND $Ca > 9.57$
		5		0.49	0.03	$Ri > 1.52043$ AND $Mg > 3.48$
		6		0.85	0.04	$Mg \leq 0.0$ AND $Ba > 0.0$
housing	SED	1	$1.5 \cdot 10^6$ $(6.0 \cdot 10^4)$	1.00 (0.91)	0.00 (0.00)	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00 (1.00)	0.00 (0.00)	$CRIM > 0.09849$ AND $TAX > 666.0$
	UG	1	$1.6 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
	EUG	1	$4.0 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
abalone	SED	1	$3.1 \cdot 10^6$ $(2.2 \cdot 10^5)$	1.00 (0.82)	0.00 (0.16)	$SEX = 2.0$ AND $SHEWEIG \leq 0.12$
		2		0.82 (1.00)	0.00 (0.00)	$LENGTH > 0.425$
	UG	1	$5.9 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		0.82	0.00	$LENGTH > 0.425$
	EUG	1	$7.4 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		0.82	0.00	$LENGTH > 0.425$
contraceptive	SED	1	$1.7 \cdot 10^7$	0.16	0.01	$WAGE \leq 37.0$ AND $CHILD > 5.0$
		2		0.86	0.00	$WAGE \leq 27.0$
		3		0.72	0.00	$WAGE > 41.0$
	UG	1	$1.3 \cdot 10^6$	0.16	0.01	$WAGE \leq 37.0$ AND $CHILD > 5.0$
		2		0.86	0.00	$WAGE \leq 27.0$
		3		0.72	0.00	$WAGE > 41.0$
	EUG	1	$1.4 \cdot 10^6$	0.16	0.01	$WAGE \leq 37.0$ AND $CHILD > 5.0$
		2		0.86	0.00	$WAGE \leq 27.0$
		3		1.00	0.11	$WAGE > 37.0$
titanic	SED	1	$6.1 \cdot 10^6$	1.00	0.00	$NPAT \leq 2.0$
		2		1.00	0.00	$NPAT > 2.0$
	UG	1	$4.2 \cdot 10^5$	0.97	0.00	$FARE \leq 79.2$
		2		0.35	0.04	$NPAT > 1.0$ AND $FARE > 79.2$
	EUG	1	$8.7 \cdot 10^5$	1.00	0.00	$CLASS \leq 3.0$
		2		-	-	$CLASS > 3.0$ AND $FARE \leq 12.875$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A13: CinterP clusters and cluster explanations for the various datasets with $\theta_1 = 1$ and $\theta_2 = 0.5$ and explanations of a maximum length of 2

Dataset	Distance	Cluster	Performance			Explanations
			Intra-homogeneity	TPR	FPR	
wine	SED	1	$1.5 \cdot 10^5$ $(5.0 \cdot 10^3)$	1.00 (1.00)	0.22 (0.00)	$\mathit{MALAC} \leq 2.68$ AND $\mathit{PROL} > 1050.0$
		2		1.00 (1.00)	0.05 (0.00)	$\mathit{PROL} \leq 740.0$
		3		0.84 (1.00)	0.31 (0.00)	$\mathit{FLAV} > 3.24$ AND $\mathit{PROL} > 740.0$
	UG	1	$1.7 \cdot 10^4$	1.00	0.00	$\mathit{HUE} > 1.24$
		2		1.00	0.00	$\mathit{PHEN} \leq 2.53$ AND $\mathit{HUE} \leq 1.24$
		3		1.00	0.00	$\mathit{PHEN} > 2.53$ AND $\mathit{HUE} \leq 1.24$
	EUG	1	$4.5 \cdot 10^4$	1.00	0.00	$\mathit{PROL} > 1265.0$
		2		1.00	0.00	$\mathit{COLINT} \leq 4.1$ AND $\mathit{PROL} \leq 1265.0$
		3		1.00	0.00	$\mathit{COLINT} > 4.1$ AND $\mathit{PROL} \leq 1265.0$
glass	SED	1	$3.8 \cdot 10^5$ $(7.8 \cdot 10^2)$	0.93 (0.80)	0.01 (0.03)	$\mathit{Fe} \leq 0.22$
		2		1.00 (1.00)	0.00 (0.00)	$\mathit{Al} \leq 0.33$ AND $\mathit{Ba} > 0.64$
		3		- (1.00)	- (0.15)	$\mathit{Al} \leq 0.33$ AND $\mathit{Ba} > 0.64$
		4		- (1.00)	- (0.01)	$\mathit{Ca} \leq 72.39$ AND $\mathit{Fe} > 0.14$
		5		- (0.92)	- (0.01)	$\mathit{Al} \leq 0.33$ AND $\mathit{Ba} > 0.64$
		6		- (0.67)	- (0.00)	$\mathit{Al} \leq 0.33$ AND $\mathit{Ba} > 0.64$
	UG	1	$5.6 \cdot 10^3$	1.00	1.00	$\mathit{Na} \leq 12.85$ AND $\mathit{Ca} > 9.02$
		2		0.97	0.08	$\mathit{Ri} \leq 1.52043$ AND $\mathit{Mg} > 2.81$
		3		0.88	0.11	$\mathit{Ri} > 1.52211$ AND $\mathit{Mg} \leq 0.0$
		4		0.85	0.69	$\mathit{Mg} \leq 0.33$ AND $\mathit{Ba} \leq 0.0$
		5		0.92	0.36	$\mathit{Si} \leq 72.39$ AND $\mathit{Ca} > 8.6$
		6		0.92	0.12	$\mathit{Na} > 13.7$ AND $\mathit{Si} > 72.79$
	EUG	1	$3.3 \cdot 10^4$	1.00	1.00	$\mathit{Na} \leq 12.85$ AND $\mathit{Ca} > 9.02$
		2		0.97	0.08	$\mathit{Ri} \leq 1.52043$ AND $\mathit{Mg} > 2.81$
		3		0.88	0.11	$\mathit{Ri} > 1.52211$ AND $\mathit{Mg} \leq 0.0$
		4		0.85	0.69	$\mathit{Mg} \leq 0.33$ AND $\mathit{Ba} \leq 0.0$
		5		0.92	0.36	$\mathit{Si} \leq 72.39$ AND $\mathit{Ca} > 8.6$
		6		0.92	0.12	$\mathit{Na} > 13.7$ AND $\mathit{Si} > 72.79$
housing	SED	1	$1.5 \cdot 10^6$ $(6.0 \cdot 10^4)$	1.00 (1.00)	0.00 (0.04)	$\mathit{RAD} \leq 8.0$ AND $\mathit{TAX} \leq 666.0$
		2		1.00 (1.00)	0.00 (0.00)	$\mathit{CRIM} > 0.09849$ AND $\mathit{TAX} > 666.0$
	UG	1	$1.6 \cdot 10^5$	1.00	0.00	$\mathit{RAD} \leq 8.0$ AND $\mathit{TAX} \leq 666.0$
		2		1.00	0.00	$\mathit{CRIM} > 0.09849$ AND $\mathit{TAX} > 666.0$
	EUG	1	$4.0 \cdot 10^5$	1.00	0.00	$\mathit{RAD} \leq 8.0$ AND $\mathit{TAX} \leq 666.0$
		2		1.00	0.00	$\mathit{CRIM} > 0.09849$ AND $\mathit{TAX} > 666.0$
abalone	SED	1	$3.1 \cdot 10^6$ $(2.6 \cdot 10^5)$	1.00 (0.95)	0.00 (0.40)	$\mathit{SEX} = 2.0$ AND $\mathit{SHEWEIG} \leq 0.12$
		2		1.00 (1.00)	0.17 (0.00)	$\mathit{LENGTH} > 0.425$
	UG	1	$5.9 \cdot 10^5$	1.00	0.00	$\mathit{SEX} = 2.0$
		2		1.00	0.17	$\mathit{LENGTH} > 0.425$
	EUG	1	$7.4 \cdot 10^5$	1.00	0.00	$\mathit{SEX} = 2.0$
		2		1.00	0.17	$\mathit{LENGTH} > 0.425$
contraceptive	SED	1	$1.7 \cdot 10^7$	1.00	0.85	$\mathit{WAGE} > 27.0$
		2		1.00	0.07	$\mathit{WAGE} \leq 29.0$ AND $\mathit{CHILD} \leq 6.0$
		3		1.00	0.11	$\mathit{WAGE} > 37.0$
	UG	1	$1.3 \cdot 10^6$	1.00	0.85	$\mathit{WAGE} > 27.0$
		2		1.00	0.07	$\mathit{WAGE} \leq 29.0$ AND $\mathit{CHILD} \leq 6.0$
		3		1.00	0.11	$\mathit{WAGE} > 37.0$
	EUG	1	$1.4 \cdot 10^6$	1.00	0.85	$\mathit{WAGE} > 27.0$
		2		1.00	0.07	$\mathit{WAGE} \leq 29.0$ AND $\mathit{CHILD} \leq 6.0$
		3		1.00	0.11	$\mathit{WAGE} > 37.0$
titanic	SED	1	$6.1 \cdot 10^6$	1.00	0.00	$\mathit{NPAT} \leq 2.0$
		2		1.00	0.00	$\mathit{NPAT} > 2.0$
	UG	1	$4.2 \cdot 10^5$	0.97	0.00	$\mathit{FARE} \leq 79.2$
		2		1.00	0.46	$\mathit{FARE} > 79.2$
	EUG	1	$8.7 \cdot 10^5$	1.00	0.00	$\mathit{CLASS} \leq 3.0$
		2		-	-	$\mathit{CLASS} > 3.0$ AND $\mathit{FARE} \leq 12.875$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

Table A14: CinterP clusters and cluster explanations for the various datasets with $\theta_1 = 2$ and $\theta_2 = 0.5$ and explanations of a maximum length of 2

Dataset	Distance	Cluster	Performance			Explanations
			Intra-homogeneity	TPR	FPR	
wine	SED	1	$1.5 \cdot 10^5$ $((5.0 \cdot 10^3))$	1.00 (1.00)	0.22 (0.00)	$MALAC \leq 2.68$ AND $PROL > 1050.0$
		2		1.00 (1.00)	0.05 (0.00)	$PROL \leq 740.0$
		3		1.00 (1.00)	0.71 (0.00)	$Mg > 88.0$ AND $PROL > 675.0$
	UG	1	$1.7 \cdot 10^4$	1.00	0.00	$HUE > 1.24$
		2		1.00	0.00	$PHEN \leq 2.53$ AND $HUE \leq 1.24$
		3		1.00	0.00	$PHEN > 2.53$ AND $HUE \leq 1.24$
	EUG	1	$5.0 \cdot 10^4$	1.00	0.00	$PROL > 1265.0$
		2		1.00	0.00	$ALCASH > 18.6$ AND $PROL \leq 1265.0$
		3		1.00	0.00	$ALCASH \leq 18.6$ AND $PROL \leq 1265.0$
glass	SED	1	$5.2 \cdot 10^4$ $(7.8 \cdot 10^2)$	1.00 (0.80)	1.00 (0.03)	$Na \leq 12.85$ AND $Ca > 9.02$
		2		0.98 (1.00)	0.15 (0.00)	$Mg > 2.81$ AND $Ca \leq 9.57$
		3		0.89 (1.00)	0.11 (0.16)	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.85 (1.00)	0.69 (0.01)	$Mg \leq 0.33$ AND $Ba \leq 0.0$
		5		1.00 (0.96)	0.64 (0.02)	$Si \leq 72.79$ AND $Ca > 8.6$
		6		1.00 (0.63)	0.31 (0.00)	$Na > 13.7$ AND $Si > 72.39$
	UG	1	$5.6 \cdot 10^3$	1.00	1.00	$Na \leq 12.85$ AND $Ca > 9.02$
		2		0.98	0.15	$Mg > 2.81$ AND $Ca \leq 9.57$
		3		0.89	0.11	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.85	0.69	$Mg \leq 0.33$ AND $Ba \leq 0.0$
		5		1.00	0.64	$Si \leq 72.79$ AND $Ca > 8.6$
		6		1.00	0.31	$Na > 13.7$ AND $Si > 72.39$
	EUG	1	$3.3 \cdot 10^4$	1.00	1.00	$Na \leq 12.85$ AND $Ca > 9.02$
		2		0.98	0.15	$Mg > 2.81$ AND $Ca \leq 9.57$
		3		0.89	0.11	$Ri > 1.52211$ AND $Mg \leq 0.0$
		4		0.85	0.69	$Mg \leq 0.33$ AND $Ba \leq 0.0$
		5		1.00	0.64	$Si \leq 72.79$ AND $Ca > 8.6$
		6		1.00	0.31	$Na > 13.7$ AND $Si > 72.39$
housing	SED	1	$1.5 \cdot 10^6$ $(6.0 \cdot 10^4)$	1.00 (1.00)	0.00 (0.04)	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00 (1.00)	0.00 (0.00)	$CRIM > 0.09849$ AND $TAX > 666.0$
	UG	1	$1.6 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
	EUG	1	$4.0 \cdot 10^5$	1.00	0.00	$RAD \leq 8.0$ AND $TAX \leq 666.0$
		2		1.00	0.00	$CRIM > 0.09849$ AND $TAX > 666.0$
abalone	SED	1	$3.1 \cdot 10^6$ $(2.6 \cdot 10^5)$	1.00 (0.98)	0.00 (0.65)	$SEX = 2.0$
		2		1.00 (1.00)	0.17 (0.00)	$LENGTH > 0.425$
	UG	1	$5.9 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		1.00	0.17	$LENGTH > 0.425$
	EUG	1	$7.4 \cdot 10^5$	1.00	0.00	$SEX = 2.0$
		2		1.00	0.17	$LENGTH > 0.425$
contraceptive	SED	1	$1.7 \cdot 10^7$	1.00	0.85	$WAGE > 27.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	UG	1	$1.3 \cdot 10^6$	1.00	0.85	$WAGE > 27.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
	EUG	1	$1.4 \cdot 10^6$	1.00	0.85	$WAGE > 27.0$
		2		1.00	0.07	$WAGE \leq 29.0$ AND $CHILD \leq 6.0$
		3		1.00	0.11	$WAGE > 37.0$
titanic	SED	1	$6.1 \cdot 10^6$	1.00	0.00	$NPAT \leq 2.0$
		2		1.00	0.00	$NPAT > 2.0$
	UG	1	$4.2 \cdot 10^5$	0.97	0.00	$FARE \leq 79.2$
		2		1.00	0.46	$FARE > 79.2$
	EUG	1	$8.7 \cdot 10^5$	1.00	0.00	$CLASS \leq 3.0$
		2		-	-	$CLASS > 3.0$ AND $FARE \leq 12.875$

Note. The values between brackets represent the results found by Carrizosa et al. (2023).

8.1 Code description

To obtain the results of this paper, four self-programmed Java classes are used that are also provided next to this paper. In alphabetical order, these classes are *BachelorThesis.java*, *IndexMinPQ.java*, *InterP.java* and *writeTXTtoCSV.java*. The results of CinterP (i.e., Tables 8, A11-A14) are obtained by running *BachelorThesis.java* with the desired (uncommented) dataset, θ_1 and θ_2. This class contains the method *readCSV* to read the dataset in. Next, the set of rules is obtained via *constructSetOfRules*, after which the compatibility set is formed with the method *explainedByRule*. The distances between observations are calculated with either *computeDissimilaritiesEuclideanSquared*, *computeDissimilaritiesUnweightedGower* or *computeDissimilaritiesExtendedUnweightedGower*, dependent on the distance measure under consideration. The method *kMeans*, together with the methods it calls, yields the k-means initialization for the MIP start, after which *solveInterP* provides initial cluster explanations. *BachelorThesis.java* extends *IndexMinPQ.java* that computes the q nearest neighbors for the extended unweighted Gower distance. Similar to CinterP, the results of InterP (i.e., Tables 7, A7-A10) are obtained by running *InterP.java* with the desired (uncommented) dataset, θ_1 and θ_2. This class does, however, not require distances between observations to be calculated since cluster allocations are already known. To obtain all datasets in CSV format, some that were obtained in TXT format had to be formatted to CSV format. Herefore, *writeTXTtoCSV.java* is used.