BMC Medical Research Methodology

RESEARCH Open Access

A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses

Pinyan Liu[1*], Han Yuan[1], Yilin Ning[1], Bibhas Chakraborty[1,2,3,4], Nan Liu[1,2,5] and Marco Aurélio Peres[1,2,6]

Abstract

Background Traditional clustering techniques are typically restricted to either continuous or categorical variables. However, most real-world clinical data are mixed type. This study aims to introduce a clustering technique specifically designed for datasets containing both continuous and categorical variables to offer better clustering compatibility, adaptability, and interpretability than other mixed type techniques.

Methods This paper proposed a modified Gower distance incorporating feature importance as weights to maintain equal contributions between continuous and categorical features. The algorithm (DAFI) was evaluated using five simulated datasets with varying proportions of important features and real-world datasets from the 2011–2014 National Health and Nutrition Examination Survey (NHANES). Effectiveness was demonstrated through comparisons with 13 clustering techniques. Clustering performance was assessed using the adjusted Rand index (ARI) for accuracy in simulation studies and the silhouette score for cohesion and separation in NHANES. Additionally, multivariable logistic regression estimated the association between periodontitis (PD) and cardiovascular diseases (CVDs), adjusting for clusters in NHANES.

Results In simulation studies, the DAFI-Gower algorithm consistently performs better than baseline methods according to the adjusted Rand index in settings investigated, especially on datasets with more redundant features. In NHANES, 3,760 people were analyzed. DAFI-Gower achieves the highest silhouette score (0.79). Four distinct clusters with diverse health profiles were identified. By incorporating feature importance, we found that cluster formations were more strongly influenced by CVD-related factors. The association between periodontitis and cardiovascular diseases, after adjusting for clusters, reveals significant insights (adjusted OR 1.95, 95% CI 1.50 to 2.55, $p=0.012$), highlighting severe periodontitis as a potential risk factor for cardiovascular diseases.

Conclusions DAFI performed better than classic clustering baselines on both simulated and real-world datasets. It effectively captures cluster characteristics by considering feature importance, which is crucial in clinical settings where many variables may be similar or irrelevant. We envisage that DAFI offers an effective solution for mixed type clustering.

Keywords Clustering, Distance measure, Feature importance, Mixed type data

*Correspondence:
Pinyan Liu
pinyanliu@u.duke.nus.edu
Full list of author information is available at the end of the article

Background

Data-driven population segmentation in clinical settings separates heterogeneous populations into homogenous groups with similar disease burdens and healthcare features [1]. Different care plans can be designed for each population subgroup using population segmentation [2]. Healthcare resource planning and evidence-based policymaking are both improved by using population segmentation analysis [3]. However, conventional segmentation has faced challenges due to the cost of patient information collection, which often involves time-consuming and labor-intensive processes such as manual data entry and extensive medical examinations. Electronic health records (EHRs) and population health surveys have become primary sources for clinical research, enhancing segmentation with their accessibility and granularity [3]. Clustering analysis is widely used for data-driven segmentation [1] but struggles with mixed type data, which includes both continuous and categorical variables. Traditional clustering algorithms, like k-means, are limited to continuous variables, posing challenges when dealing with mixed type data [4].

Clustering mixed type data in biomedicine holds significant value due to its ability to address complex and heterogeneous data sets [5]. It enables patient stratification [6], pattern discovery in genomics [7], personalized medicine [8], disease marker identification [9], and linking genetics with brain imaging [10]. Another important benefit of clustering is to help gain different insights into associations among people sharing similar baseline characteristics [11].

Analysts have attempted to transform mixed type data into a uniform type for clustering [12, 13], for example, converting all variables into continuous ones and using k-means to calculate Euclidean distances [14, 15]. But this transformation often results in a loss of information, impairing segmentation quality. The k-prototypes algorithm [16], an adaptation of k-means, combines squared Euclidean and matching distances to cluster mixed data. User-defined weighting factors set the importance of each variable type, echoing the limitations of using Gower's distance. However, most previous weighting methods don't fully address the challenges of mixed data [17, 18], often failing to balance the influence of continuous and categorical variables effectively [19, 20].

Model-based clustering assumes data comes from various probability distributions, with each representing a different cluster [21–23]. This method is effective for mixed-type datasets where other clustering techniques struggle [24]. It uses distributions (like Gaussian or multinomial) to model clusters and the Expectation-Maximization (EM) algorithm to estimate distribution parameters and assign data points to clusters [25]. While adaptable and useful for determining cluster membership probabilities and the number of clusters, it can falter if its assumptions are not met.

KAMILA [26] (KAymeans for MIxed Large data), an efficient variant of the k-means algorithm [15], addresses the challenge of clustering mixed continuous and categorical data without heavy parametric assumptions [23]. Simulation results indicate that it handles large datasets effectively, addressing both data types without explicit weighting, with potential advantages over methods such as normal-multinomial mixture models and the Modha-Spangler weighting approach in certain contexts [27, 28].

In addition to clustering algorithms, many distance metrics were proposed for mixed type data, which integrate different types of variables' distance, with Gower's distance being the most popular [29]. However, Gower distance can be dominated by categorical variables, neglecting the important differences within variable types [30]. This paper explores the efficacy of current clustering techniques for handling mixed type data and identifies a gap in using Gower distance to consider the influence of both continuous and categorical variables evenly.

To address this gap, we introduce an innovative two-step framework for mixed type clustering with a new modified and weighted Gower distance, DAFI-Gower, considering distance adjustment and feature importance to improve clustering quality and interpretability. The performance of the DAFI-Gower method was evaluated using simulated data and compared with other mixed type clustering techniques to assess clustering quality. Additionally, its application was demonstrated using real-world datasets to identify distinct health profiles, particularly those related to periodontitis (PD) and cardiovascular diseases (CVDs). The method also aids in adjusting for confounders in association analyses through clustering.

Methods

This section details our proposed two-step framework: first, a distance matrix is constructed to quantify a balanced distance measurement between different types of variables; second, a clustering algorithm is applied using this distance matrix. The distance matrix employs our modified Gower distance, incorporating distance adjustment and mutual information-based feature importance (DAFI) weights.

Gower distance

Distance metrics created for mixed data sets, such as Gower distance [29], are widely used for measuring the degree of dissimilarity between observations when mixed type features are present. Assume a mixed type dataset

$\boldsymbol{X}$ with n observations has p features, where the first h features are continuous, and the remaining features from $h+1$ to p are categorical. Therefore, the Gower distance between two observations $\boldsymbol{x_i} = (x_{1i}, x_{2i}, ., x_{ji}, ., x_{pi})$ and $\boldsymbol{x_k} = (x_{1k}, x_{2k}, ., x_{jk}, ., x_{pk})$ from dataset X $(i, k \in \{1,2,\ldots,n\})$ is:

$$d(\boldsymbol{x_i}, \boldsymbol{x_k}) = \frac{1}{p}\left\{\sum{}_{j=1}^{p} d_j\left(x_{ji}, x_{jk}\right)\right\},$$

where $d_j\left(x_{ji}, x_{jk}\right)$ is defined differently for continuous and categorical variables:

$$d_j\left(x_{ji}, x_{jk}\right) = \begin{cases} \frac{|x_{ji}-x_{jk}|}{R_j} & \text{if } j \in \{1,2,\ldots,h\} \\ I\left(x_{ji} \neq x_{jk}\right) & \text{if } j \in \{h+1, h+2,\ldots,p\}, \end{cases}$$

where R_j is the range for the value of j^{th} continuous feature, with x_{ji} and x_{jk} being the values of the j^{th} feature for observations $\boldsymbol{x_i}$ and $\boldsymbol{x_k}$ separately. $I(x_{ji} \neq x_{jk})$ is 1 if $x_{ji} \neq x_{jk}$ and 0 otherwise. The default setting in Gower distance specifies equal weights for all variables, while a vector of variable importance could be applied to $d_j(\boldsymbol{x_i}, \boldsymbol{x_k})$ as an optional choice. Gower distance in a clinical context compares patients by assigning higher weights to more critical variables, such as focusing on recovery time and complications for evaluating treatment outcomes [31], ensuring that these attributes have a greater influence on the similarity calculation for more accurate patient comparisons.

In conventional Gower distance calculation, categorical variables often impact the results more than continuous ones because a categorical difference can easily reach the maximum distance of 1. In contrast, the difference in a continuous variable only reaches 1 when calculated between the extremes (minimum and maximum values). This discrepancy can cause categorical variables to disproportionately influence the distance metric and affect subsequent analysis, as categorical differences are more common and can overshadow the nuanced differences captured by continuous variables. For instance, two individuals with a 63-year age difference (continuous) may have a smaller Gower distance than those differing only by sex (categorical).

To address the imbalance in Gower distance calculations, D'Orazio [30] suggested using the inter-quartile range (IQR) for continuous variables to reduce outlier impacts and balance variable contributions but not having similar adjustments made for categorical variables.

Proposed modified Gower distance

To improve the handling of categorical variables, we converted categorical values into dummy variables to refine the distance measure, aligning it with continuous variables (Fig. 1). However, this method can disproportionately affect datasets with variables that have many categorical levels. To address this, adjusted weights were proposed for categorical features using continuous features as references, and for continuous features, scaling the Manhattan distance (the absolute difference) by the IQR is kept using as recommended by D'Orazio [30]. This method achieves a more fine-grind and balanced assessment of dissimilarity for mixed type variables, ensuring continuous and categorical features contribute equitably to the overall distance measure. Below are the details for calculating weights.

Continuous features

For each continuous feature, the sum of the absolute differences is calculated between all possible pairs of observations. Mathematically, for each feature column $X_j (j \in \{1,2,\ldots,h\})$, the sum of distances is given by:

$$D_{X_j} = \sum{}_{i=1}^{n} \sum{}_{k=1}^{n} \left|X_{ji} - X_{jk}\right|$$

where n is the number of observations. Then, the calculated sum distance is divided by the difference between the first and third quartiles [30]. Mathematically:

$$Q_{X_j} = \frac{D_{X_j}}{Q_3\left(X_j\right) - Q_1\left(X_j\right)}$$

where $\mathrm{Q}_3\left(X_j\right)$ and $Q_1\left(X_j\right)$ are the third and first quartiles of the j^{th} continuous feature.

Categorical features

Each categorical feature $X_j, (j \in \{h+1, h+2,\ldots,p\})$, with L levels is converted into L dummy variables first, where $X_{j[m]}$ denotes the m^{th} dummy variable $(m \in \{1,2,\ldots,L\})$. Then, the sum of the differences is calculated for all L dummy variables between each pair of observations. Mathematically, for each feature column X_j, the sum of distances is given by:

$$D_{X_j} = \left\langle \sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{m=1}^{L}\left\{ I\left(x_{(j[m])i} \neq x\left(j[m]\right)k\right)\right\}\right\rangle / L$$

where $x_{(j[m])i}$ is the m^{th} dummy variable of categorical feature $X_j$for the i^{th} observation, and L is the total number of dummy variables for the categorical feature X_j. Then, the distance is averaged by the number of dummy variables L. An example for comprehension is available in Supplementary Fig. 1.

As mentioned earlier, the adjusted distance for categorical variables is calculated then. This is done by scaling the original distance D_{X_j} for categorical variables. We multiply by the scaling factor (α_j), which is the ratio of the average of normalized distances for the continuous

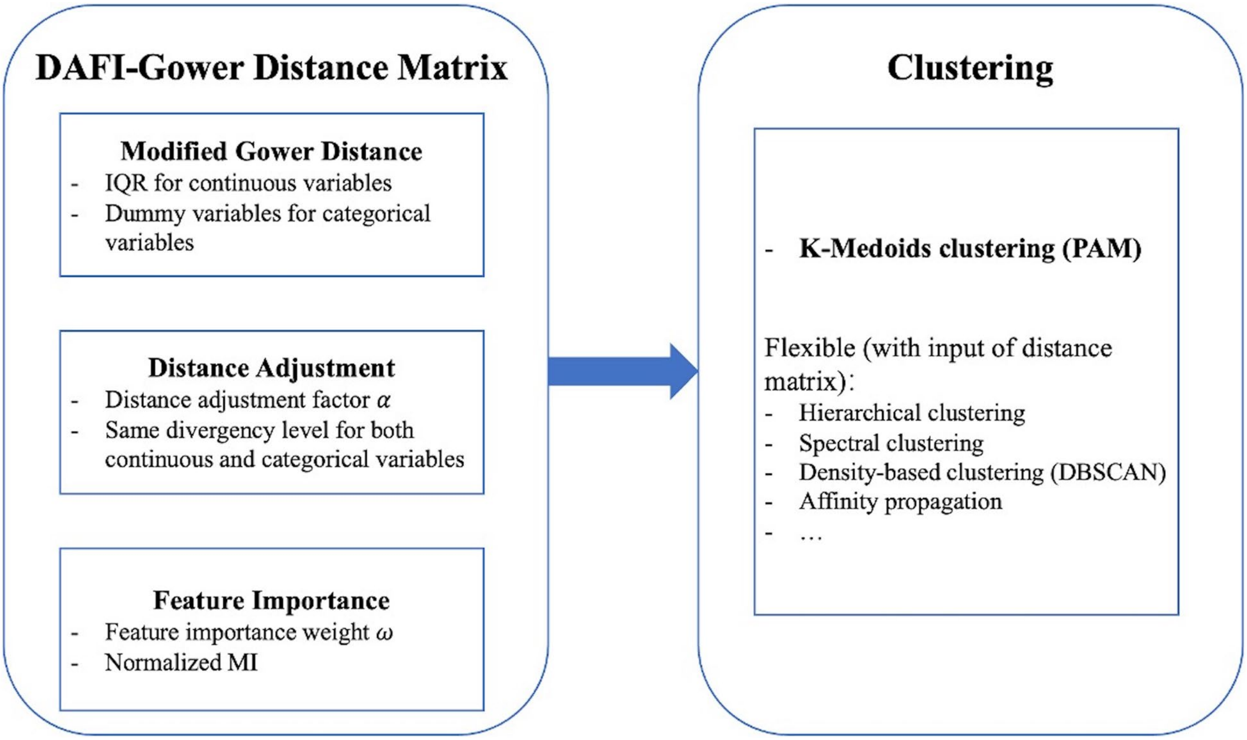
DAFI-Gower Distance Matrix
Modified Gower Distance
- IQR for continuous variables
- Dummy variables for categorical variables
Distance Adjustment
- Distance adjustment factor α
- Same divergency level for both continuous and categorical variables
Feature Importance
- Feature importance weight ω
- Normalized MI
Clustering
- K-Medoids clustering (PAM)
Flexible (with input of distance matrix):
- Hierarchical clustering
- Spectral clustering
- Density-based clustering (DBSCAN)
- Affinity propagation
- …

Fig. 1 The overall framework of the DAFI-Gower algorithm used for mixed type clustering

features $Q_{X_q}(q \in \{1,2,\ldots,h\})$ to the categorical features $D_{X_j}(j \in \{h+1, h+2,\ldots,p\})$:

$$\alpha_j = \frac{\frac{1}{h}\sum_{q=1}^{h} Q_{X_q}}{D_{X_j}} = \frac{\overline{Q}_X}{D_{X_j}}$$

As previously mentioned, Q_{X_q} and D_{X_j} represent total distance contributed by a continuous feature X_q and a categorical feature X_j, respectively, they can be interpreted as the contributions to the overall distance. Therefore, with the scaling factor α_j in the modified Gower distance to adjust the $D\prime_{X_j} = \alpha_j x D_{X_j} = \overline{Q}_X$, we guarantee the balanced contribution of continuous and categorical features toward the total distance.

Proposed modified and weighted Gower distance with feature importance

After aligning the distance measures for continuous and categorical features, the modified Gower distance matrix is used for clustering. Clustering quality and interpretations are critical, especially in clinical settings. Therefore, identifying key features is crucial for effective analysis, moving beyond mere grouping subjects to determining which features significantly influence group formation. This improves clustering interpretations by addressing the limitations of its unsupervised nature. Therefore, incorporating feature importance as weights was proposed in the modified Gower distance to obtain the final dissimilarity matrix.

Information theory concepts like entropy and mutual information (MI) are crucial for evaluating feature importance [32]. Before the calculation of entropy and MI, continuous variables are converted into categorical variables based on quartiles [19, 33]. Entropy ($H(X)$) measures the uncertainty of a discrete variable, indicating higher information content with greater unpredictability [34]. Furthermore, MI ($I(X_1; X_2)$) measures how much knowing one variable reduces the uncertainty of another, which is crucial for selecting relevant features by quantifying the information shared between variables. However, MI can suffer from redundancy, capturing overlapping information from multiple variables and overestimating their importance [35]. Normalized MI (NMI), a variant of MI, addresses this by providing better accuracy and handling redundancy [36]. It is defined as:

$$NMI(X_1; X_2) = \frac{I(X_1; X_2)}{max\{H(X_1), H(X_2)\}}$$

Horibe [37] demonstrated that MI scaled by the maximum of entropy is a valid distance metric, making it a normalized similarity metric [38]. Mousavi and Elahe [39] also introduced a unified framework based on mutual information control variables' contribution to distance measurement, which could prevent unnecessary information. Therefore, to measure feature importance, the average NMI of each feature with all other features was proposed as the weighting parameter ω. Procedures for calculating the weighting parameter ω for the j^{th} feature are shown below (Algorithm 1). In detail, for each feature *j*, the algorithm calculates how much information it shares with every other feature using NMI. The total NMI score for feature *j* is then averaged and stored in a vector. Finally, the importance weights are normalized by dividing the sum of all weights, ensuring they add up to 1.

Algorithm 1 Pseudo-code for feature importance weights

```
  $\mathbf{NMI} = \mathbf{0}_{\text{ncol}(X)}$
  for $j = 1$ to $\text{ncol}(X)$ do
      $X_{*j} = X[\ ,\text{-}j]$
      $X_j = X[\ ,j]$
      $Weight_j = 0$
      for $m = 1$ to $\text{ncol}(X_{*j})$ do
          $Weight_j = Weight_j + \text{NMI}\ (\text{X}_j, X_{*j}[\ ,m])$
      end for
      $Weight_j = Weight_j / \text{ncol}(X_{*j})$
      $\mathbf{NMI}[j] = Weight_j$
  end for
  $\omega = \mathbf{NMI}/\text{sum}(\mathbf{NMI})$
return $\omega$
```

This approach considers both the intrinsic uncertainty of features and their mutual dependency on each variable. Therefore, the entire DAFI-Gower distance is calculated as:

$$DAFI - Gower\ Distance = DAFI(\boldsymbol{x_i}, \boldsymbol{x_k}) = \frac{1}{p}\left\{\sum{}_{j=1}^{p} \alpha_j * \omega_j * d_j\left(x_{ji}, x_{jk}\right)\right\},$$

where $d_j\left(x_{ji}, x_{jk}\right)$ is calculated through:

$$d_j\left(x_{ji}, x_{jk}\right) = \begin{cases} \frac{\sum_{i=1}^{n}\sum_{k=1}^{n}|x_{ji}-x_{jk}|}{IQR_j} & \text{if } j \in \{1,2,\ldots,h\} \\ \frac{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{m=1}^{L}\left\{I\left(x_{(j[m])i} \neq x_{(j[m])k}\right)\right\}}{L} & \text{if } j \in \{h+1, h+2,\ldots,p\}, \end{cases}$$

According to the formula for the DAFI-Gower distance, which involves summation terms, the calculation requires iterating over i (from 1 to n) for each k (also ranging from 1 to n). Therefore, for any given *L*, the time complexity is $O(n^2)$, indicating that as the size of the dataset (n) increases, the number of computations grows quadratically. Analysis was also conducted to compare the execution time of DAFI-Gower algorithm using NHANES.

The proposed study design using DAFI-Gower algorithm

The novel mixed type clustering technique (Fig. 1) provides a balanced distance measurement and incorporates feature importance as weights. In the first phase, the DAFI-Gower algorithm is designed to calculate the distance matrix through three main parts: calculating the modified Gower distance, adjusting the distance with factor α, and incorporating feature importance ω as weights. The resulting DAFI-Gower distance matrix is then used for subsequent clustering analysis. Partitioning around medoids (PAM) [40, 41]) is chosen, as it allows clustering based on a predefined distance matrix. While our current framework employs PAM, it can accommodate any clustering algorithm that accepts a predefined distance matrix.

Simulation settings

Simulation studies were conducted to compare the accuracy of the proposed method against commonly used clustering techniques for mixed type data, evaluate improvements from different feature importance strategies, and assess scalability with heterogeneous datasets with different proportions of important features contributing to clustering (feature importance). The adjusted Rand index (ARI) was calculated to measure the similarity between true and predicted clustering [42], where ARI ranges from -1 (indicating totally different) to 1 (indicating a perfect match), with 0 indicating the similarity between the two partitions is what would be expected by chance [42].

Five simulation settings with different proportions of feature importance were generated (Supplementary Table 1), reflecting real-world clinical scenarios with many redundant variables. Three underlying true clusters with cluster sizes of 60, 60, and 80 were assumed. Each study had five continuous and five categorical variables. Five continuous variables were drawn from normal distributions with cluster-specific means and standard deviations. If the variable has the same mean for all clusters, it did not contribute to cluster formation. Additionally, five categorical variables were generated with predefined categories and cluster-specific probabilities. Similarly, if the variable has the same probabilities for all clusters, it had no contribution to cluster formation. Data points were assigned to clusters based on these parameters, resulting in distinct cluster profiles. This process was repeated 500 times to create datasets for further analysis. To validate the reproducibility of DAFI-Gower, we conducted additional simulations with cluster sizes of 30, 60, and 90, simulating scenarios with less balanced cluster sizes. All other parameters were kept consistent with the previous settings. The scenarios were:

(1) 4/5 of both continuous and categorical variables contribute to clustering.
(2) 3/5 of both continuous and categorical variables contribute.
(3) 2/5 of both continuous and categorical variables contribute.
(4) 1/5 of both continuous and categorical variables contribute.
(5) 2/5 of continuous and 4/5 of categorical variables contribute.

Simulation scenario 5 was designed to reflect real-world settings like social science and EHRs, where categorical variables usually dominate the datasets. For each setting, 500 datasets were generated. After using DAFI-Gower to calculate the distance matrix, we applied PAM for the clustering analysis.

In order to compare the overall performance of DAFI-Gower with commonly used mixed type clustering techniques, as well as evaluate each component's importance in DAFI-Gower, we designed simulation studies into three parts:

Part I: Comparison of DAFI-Gower with commonly used mixed type clustering techniques:

(1) K-prototypes [16] (an adaptation of k-means, combines squared Euclidean and matching distances to cluster mixed data). This method was implemented by its authors in the *clustMixType* R package [43];
(2) K-means (categorical factors are treated as numeric; non-numeric categories are assigned arbitrary numeric codes for clustering);
(3) K-modes (discretize all continuous variables into categorical ones based on quartiles; for example, given a continuous variable x, define the quartile boundaries Q1 (25th percentile of x), Q2 (50th percentile of x), Q3 (75th percentile of x). Now, define a new categorical variable $x_{cat} = 1$ ($if\ x \leq Q1$); 2 ($if\ Q1 \leq x \leq Q2$); 3 ($if\ Q2 \leq x \leq Q3$); 4 ($if\ x > Q3$).)
(4) K-prototypes with Gower distance;
(5) KAymeans for MIxed Large data(KAMILA) [44] This clustering algorithm integrates the k-means algorithm and Gaussian-multinomial mixture models. Like k-means, avoids making strict parametric assumptions about continuous variables. Similar to Gaussian-multinomial mixture models, KAMILA effectively balances the contributions of continuous and categorical variables without presetting weights, instead relying on a data-driven density estimator [23];
(6) PAM with Gower distance.

Part II: Comparison of DAFI-Gower with three feature importance algorithms as baseline feature importance methods, to evaluate the performance of NMI (see Table 1 elaborating on the properties of these methods):

(1) Distance-based on co-occurrence of values (Algo_distance) [19, 33]. It discretized the continuous attributes and calculated the distance between every pair of attribute values for all attributes. Then, they modified the k-means algorithm to contain the distribution of all categorical values in a cluster.
(2) Two-stage approach (FR & FS) [45]. This methodology employs three fundamental strategies: filter, wrapper, and hybrid (a fusion of both filter and wrapper) to identify relevant and non-redundant features. This work introduces a robust and efficient two-phase (i.e., feature ranking and feature selection) method. In the stage of feature ranking, they employed entropy and mutual information to provide a normalization to rank features, and then define the feature weight as the degree of the unique or non-shared information (entropy minus the normalization value) of an individual feature.
(3) Standard mutual information (elaborated in Sect. 2.3). This measures the amount of information shared between two random variables, quantifying

Table 1 Comparison of different components of three baseline feature importance algorithms and different combinations of DAFI-Gower algorithm

Methods		**Distance Measurement**			**Feature Importance**
		Quantile scale	**Dummy conversion**	**Balanced "Quantile +Dummy"**	**Mutual Information**
Baseline feature importance method 1	1.Algo_distance+feature importance 2.PAM + Gower distance	×	×	✓	✓
Baseline feature importance method 2	1.Two-stage feature importance 2.PAM + Gower distance	×	×	×	✓
Baseline feature importance method 3	1.Standard mutual information 2.PAM + Gower distance	×	×	×	✓
Proposed new method 1	PAM + modified Gower distance	✓	✓	×	×
Proposed new method 2	1.Distance adjustment 2.PAM + modified Gower distance	✓	✓	✓	×
Proposed new method 3	1.Normalized MI + feature importance 2.PAM + modified Gower distance	✓	✓	×	✓
DAFI-Gower Distance with PAM	1.Distance adjustment + Normalized MI 2. PAM + modified Gower distance	✓	✓	✓	✓

how much knowing one variable reduces uncertainty about the other.

Part III: Comparison of different DAFI-Gower combinations (Fig. 1) to evaluate the individual components of the technique (see Table 1 elaborating on the properties of these methods):

(1) Distance was calculated using modified Gower distance, and clustering was performed using PAM.
(2) Distance was calculated using modified Gower distance and distance adjustment, and clustering was performed using PAM.
(3) Distance was calculated using modified Gower distance and feature importance, and clustering was performed using PAM.
(4) DAFI-Gower distance matrix was calculated using all three components (modified Gower distance, distance adjustment, and feature importance), and clustering was performed using PAM.

The median ARI values, along with the 25th and 75th quartile, were reported for all statistics. All data analysis was implemented in R. All algorithms were performed with default values, except for the specified data and the number of clusters. For k-means clustering, the *'nstart'* parameter was set to 10 rather than the default of 1 to improve result stability. The source code of the data analysis is available from the authors in:[https://github.com/Pinyan-Liu/DAFI-Gower-Distance] [46].

Empirical study settings

To demonstrate the practical implementation of the DAFI-Gower distance clustering algorithm, we utilized real datasets to evaluate the proposed method's performance against the methods in the three parts described above, following the study design in Fig. 1.

CVDs are inflammatory conditions of the coronary arteries and the leading cause of death globally. In recent decades, investigators have focused on the impact of PD on CVDs, a potential risk factor, promoting the development and instability of arterial atheroma [47]. To deliver a comprehensive representation of the relationship, we proposed to estimate the association between PD severity following the 2012 Centers for Disease Control and Prevention (CDC)/American Academy of Periodontology (AAP) [48] (see Supplementary Table 5) and CVDs by adjusting for distinct clusters with similar characteristics. We used data from the 2011–2014 National Health and Nutrition Examination Study (NHANES) [49, 50], a continuous study conducted by the Centers for Disease Control and Prevention (CDC) to assess the health and nutrition status of the US population who are not in institutions. This study collects large amounts of quantitative and qualitative data using face-to-face interviews, physical evaluations, computerized questionnaires, and laboratory analyses [50] and was conducted following the Strengthening the Reporting of Observational Studies in Epidemiology guidelines [51].

People over 30 were studied due to the availability of data on PD. Information on PD was clinically obtained and available in NHANES. The analysis excluded participants without complete periodontal exams or complete covariate data. We defined the indicator of CVDs as being diagnosed with either congestive heart failure or stroke [52]. Information on CVDs was self-reported and available in NHANES. The data set included 3,760 observations with complete data. Sixteen variables (see Supplementary Table 2) were used to generate clusters, including demographic variables, clinical oral health data, and CVD-related variables. The number of clusters was predetermined to be four [48]. Clustering analysis used the same 13 approaches as the simulation study (Parts I, II, and III). Clustering quality was assessed using descriptive statistics and Silhouette scores measuring cohesion and separation [53]. Although Silhouette scores should be interpreted cautiously due to their dependence on the selected distance metric and lack of comparability across different distances, it is also a well-used tool for descriptive analysis or as one of multiple metrics in assessing clustering outcomes [54]. Multivariable logistic regression was used to estimate the association between PD and CVDs, adjusting for clusters generated in the previous step. Additionally, feature significance calculated using NMI described in Sect. 2.3 was plotted to show clustering contribution.

Results

Simulation results

Results on different baseline mixed type clustering (part I)

Figure 2a and Supplementary Table 3 show the median ARI index together with the 25th and 75th percentiles on five mixed type datasets. From the results, the DAFI-Gower distance with PAM performed consistently the best, especially when only a small portion of variables contribute to the clustering formation. Although KAMILA performed as well as DAFI-Gower in simulations 1 and 5, it does not provide feature importance, which is significant in our study settings.

Results on different feature importance algorithms (part II)

Figure 2b and Supplementary Table 4 (blue background) show the median ARI index together with the 25th and 75th percentiles on five mixed type datasets. Separate use of distance adjustment or feature importance performed relatively well in specific settings, but none showed consistent performance.

Results on different parts of the proposed method (part III)

Figure 2b and Supplementary Table 4 (the yellow part) compare DAFI-Gower distance and PAM clustering to show each part's improvement. The results indicate that both the DAFI-Gower distance and its component (just the distance adjustment factor or only feature importance weights) performed well. When the data included more non-contributable features (simulation 5), DAFI-Gower distance with PAM performed better than all other approaches.

From above, techniques on mixed type data are often better than those on quantitative or categorized data (Table 1). Our novel approaches for mixed type variables performed better than existing methods in settings investigated, especially when employing the DAFI-Gower distance with PAM on datasets with more irrelevant features. Our approaches also give feature importance, an advantage over traditional mixed type clustering algorithms for visualization and real-world application.

Results from the additional analysis with cluster sizes of 30, 60, and 90 are shown in Supplementary Tables 7 and 8. The outcomes are consistent with previous findings. DAFI-Gower performs better than all other methods in both the baseline mixed type clustering comparison (Part I) and feature importance algorithms (Part II). In Part III, which compares DAFI-Gower to its individual components, DAFI-Gower ranks first in simulations 8, 9, and 10—where more irrelevant variables are included—and second in the remaining cases.

Empirical study results

Clustering results

Figure 3 presents the silhouette index according to approaches. The new DAFI-Gower distance and PAM clustering approach has the greatest Silhouette score (0.79), suggesting the best cluster cohesion and separation. The baseline mixed type clustering approaches performed worse than feature-importance clustering. The green area of Fig. 3 demonstrated that distance adjustments and feature importance weights improved clustering performance, but the combination did more.

Descriptive statistics for four clusters and the overall sample were presented in Table 2. Four distinct segments were identified. Cluster 1 is the healthier and younger, with a mean age of 50.1 ± 13.8 years, and has the lowest proportion of obese people (35.3%), 59.9% of non-smokers, and 86.3% non-diabetes. For PD, Cluster 1 also has the highest proportion of healthy people (61.4%) and non-severe PD people (92.8%). Another significant characteristic is that Cluster 1 can be treated as "the cluster with no CVDs" with only 1.8% of participants having CVDs. In contrast, Cluster 3 was the least healthy and oldest, with a mean age of 61.5 (SD: 12.3) years; 77.3% of participants were overweight or obese, only 37.9% were non-smokers, and 62.1% were non-diabetic. For PD, Cluster 3 also has more non-severe PD people (13.6%) and is "the cluster with the highest probability of CVDs" with 34.8% of participants having CVDs. Cluster 2 and 4 are the relatively intermediate ones.

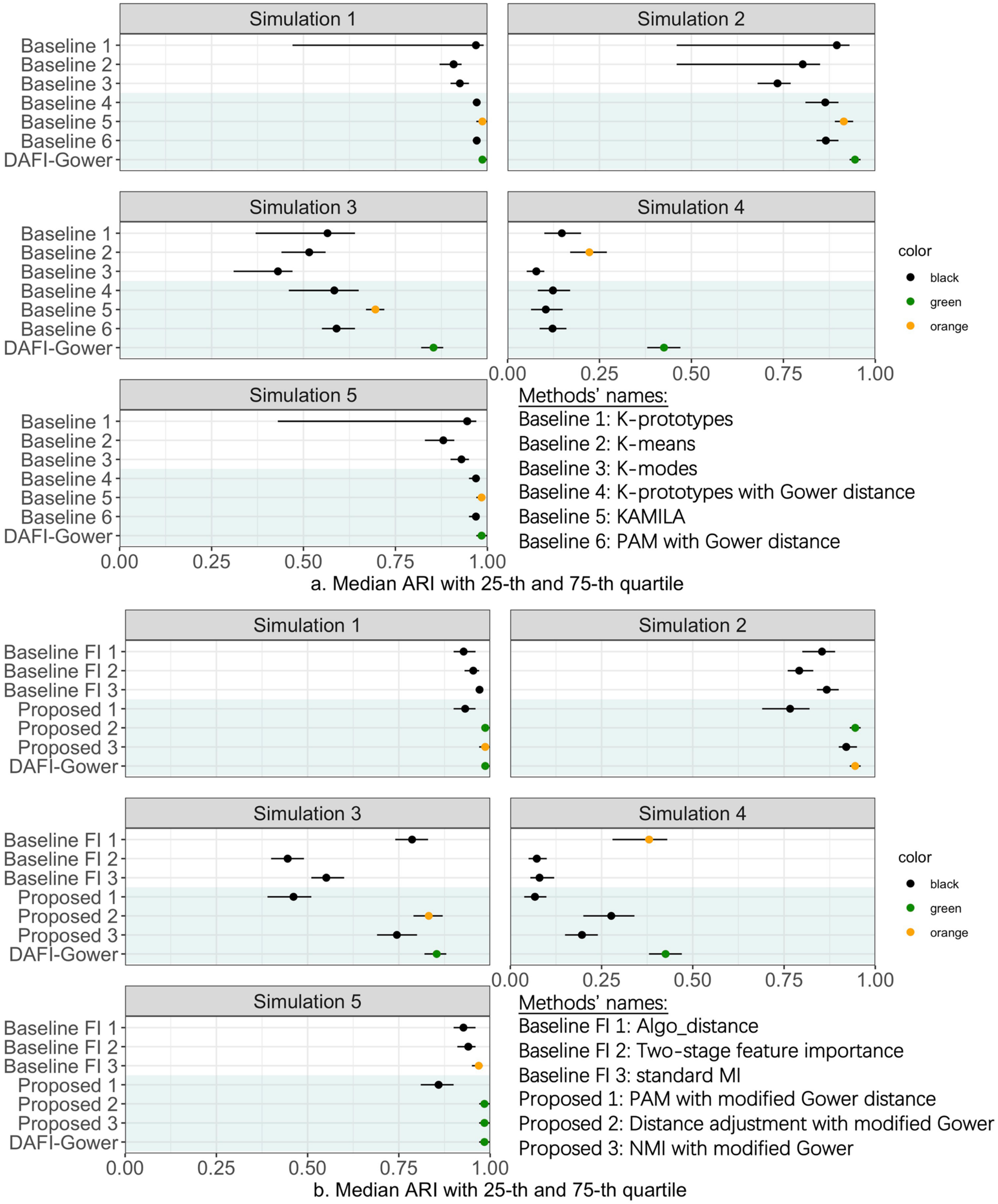

Fig. 2 Simulation results of median ARI with 25th and 75th quartile. The first and second ranks are indicated by green and orange boldface, respectively. **a** Results on comparing different baseline mixed type clustering techniques (Part I). **b** Results on comparing different baseline feature importance techniques (Part II) and the comparison of different DAFI-Gower combinations (Part III)

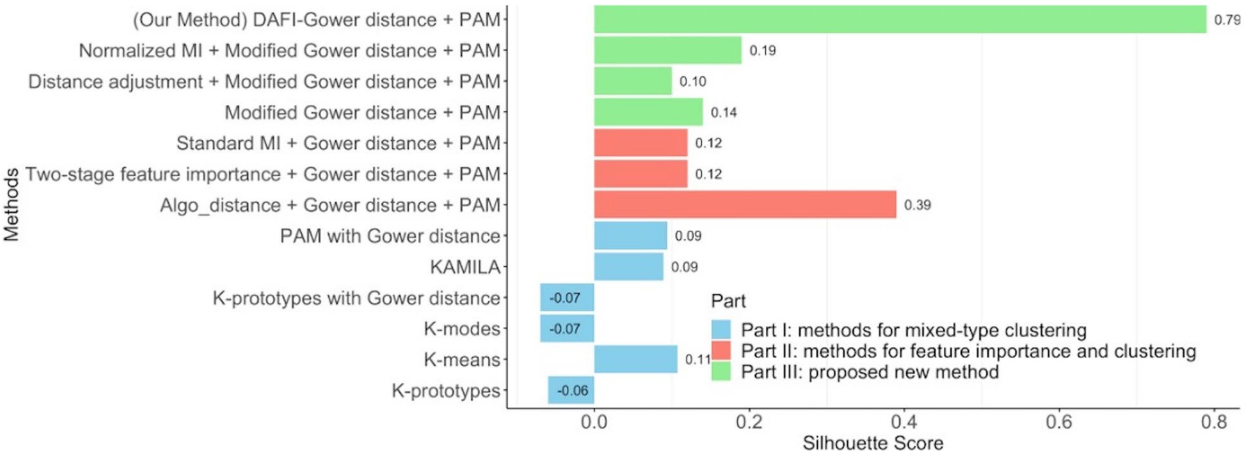

Fig. 3 Comparative analysis of clustering performance using the silhouette index across baseline and novel methods on NHANES data. The graph delineates method categories through color differentiation

Figure 4 shows clustering-related feature contributions to help assess cluster characteristics. From the figure, coronary heart disease, angina pectoris, heart attack, and emphysema are among the averaged contributions, which are all CVD-related. Cluster 1's low CVD prevalence may be explained by the top four features.

Association between PD and CVDs

To investigate the benefits of using clusters in analyzing the association between PD severity (categorized as non-severe: combining none, mild, and moderate stages; and severe stage) and CVD cases, we conducted three logistic regression analyses: Model 1 - unadjusted analysis; Model 2 - conventional adjusted analysis (that adjusted for individual characteristics); Model 3 - proposed adjusted analysis (that adjusted for cluster membership instead of individual characteristics). From Table 3, we found that the association between PD and CVDs reveals significant insights in Model 1 (odds ratio (OR) 1.97, 95% confidence interval (CI) 1.55 to 2.49, $p=0.0045$) but not in Model 2, which indicates that adjusted for too many confounders may diminish the observed association supported by domain knowledge and previous research [47]. However, this issue can be resolved by adjusting for cluster membership, which provides a more accurate representation of the data indicated by Model 3 (adjusted OR 1.95, 95% CI 1.50 to 2.55, $p=0.012$). Patients diagnosed with severe PD had approximately twice the odds of having CVDs compared to those diagnosed with no PD or non-severe PD.

In addition, after adjusting for clustering, the analysis reveals significantly higher CVDs risks in clusters 3 and 4, suggesting that these groups warrant further investigation to identify high-risk populations for targeted interventions in future studies, which is also consistent with the descriptive statistics presented in Table 3 that clusters 3 and 4 have the significantly higher proportions of CVDs cases. This phenomenon suggests that the clusters identified by the new method may encompass inherent characteristics, possibly undetected or unmeasured risk factors, strongly associated with CVDs risk.

DAFI-Gower algorithm execution time

Supplementary Table 9, includes the computational cost of DAFI-Gower. The execution times of different algorithms on NHANES, with 3,760 observations and 18 variables, are summarized.

Discussion

In this article, we developed the DAFI-Gower technique, an innovative two-step framework for mixed type clustering to mitigate the limitations of single-type methods and unbalanced measurements like the Gower distance. The novelty of DAFI-Gower lies in its distance adjustment, allowing the same level of contributions for different feature types. Additionally, the subsequent inherent feature importance weights, which account for contributions to clustering, enhance the interpretability of the clustering analysis, making it more informative. Results from both simulation and real-world studies demonstrate that the DAFI-Gower distance with PAM clustering is a flexible approach. In real-world studies, DAFI-Gower not only performed well in clustering patients but also helped to improve association studies.

The DAFI-Gower mitigates some limitations of mixed type clustering. Previous studies have often treated continuous and categorical data separately or applied simplistic methods that fail to capture the nuanced

Table 2 Descriptive statistics table for four clusters and the overall sample for DAFI-Gower clustering results

	Cluster 1 (*N* = 2809)	**Cluster 2 (*N* = 779)**	**Cluster 3 (*N* = 66)**	**Cluster 4 (*N* = 106)**	**Overall (*N* = 3760)**
Age					
Mean (SD)	50.1 (13.8)	47.8 (12.3)	64.9 (12.3)	61.5 (14.1)	50.2 (13.8)
Gender					
Male	1296 (46.1%)	596 (76.5%)	41 (62.1%)	69 (65.1%)	2002 (53.2%)
Female	1513 (53.9%)	183 (23.5%)	25 (37.9%)	37 (34.9%)	1758 (46.8%)
Age categories					
30–39 Years	769 (27.4%)	237 (30.4%)	3 (4.5%)	9 (8.5%)	1018 (27.1%)
40–49 Years	696 (24.8%)	208 (26.7%)	4 (6.1%)	17 (16.0%)	925 (24.6%)
50–59 Years	584 (20.8%)	181 (23.2%)	10 (15.2%)	14 (13.2%)	789 (21.0%)
60–69 Years	458 (16.3%)	115 (14.8%)	23 (34.8%)	33 (31.1%)	629 (16.7%)
70–80 Years	302 (10.8%)	38 (4.9%)	26 (39.4%)	33 (31.1%)	399 (10.6%)
Race/Hispanic origin					
Mexican American	242 (8.6%)	162 (20.8%)	4 (6.1%)	10 (9.4%)	418 (11.1%)
Other Hispanic	226 (8.0%)	82 (10.5%)	4 (6.1%)	10 (9.4%)	322 (8.6%)
Non-Hispanic White	1406 (50.1%)	285 (36.6%)	45 (68.2%)	46 (43.4%)	1782 (47.4%)
Non-Hispanic Black	531 (18.9%)	189 (24.3%)	11 (16.7%)	26 (24.5%)	757 (20.1%)
Other Race-Including Multi Racial	404 (14.4%)	61 (7.8%)	2 (3.0%)	14 (13.2%)	481 (12.8%)
Education level					
Less than 9th grade	81 (2.9%)	76 (9.8%)	4 (6.1%)	12 (11.3%)	173 (4.6%)
9-11th grade (includes 12th grade with no diploma	232 (8.3%)	158 (20.3%)	6 (9.1%)	14 (13.2%)	410 (10.9%)
High school graduate/GED or equivalent	499 (17.8%)	231 (29.7%)	14 (21.2%)	26 (24.5%)	770 (20.5%)
Some college or AA degree	862 (30.7%)	235 (30.2%)	22 (33.3%)	30 (28.3%)	1149 (30.6%)
College graduate or above	1135 (40.4%)	79 (10.1%)	20 (30.3%)	24 (22.6%)	1258 (33.5%)
BMI categories					
Healthy	804 (28.6%)	216 (27.7%)	15 (22.7%)	26 (24.5%)	1061 (28.2%)
Overweight	1014 (36.1%)	249 (32.0%)	18 (27.3%)	35 (33.0%)	1316 (35.0%)
Obese	991 (35.3%)	314 (40.3%)	33 (50.0%)	45 (42.5%)	1383 (36.8%)
Smoking status					
Never smoke	1682 (59.9%)	206 (26.4%)	25 (37.9%)	45 (42.5%)	1958 (52.1%)
Former smoker	738 (26.3%)	206 (26.4%)	27 (40.9%)	38 (35.8%)	1009 (26.8%)
Current smoker	389 (13.8%)	367 (47.1%)	14 (21.2%)	23 (21.7%)	793 (21.1%)
Diabetes					
Non-diabetes	2423 (86.3%)	619 (79.5%)	41 (62.1%)	76 (71.7%)	3159 (84.0%)
Prediabetes	146 (5.2%)	42 (5.4%)	3 (4.5%)	3 (2.8%)	194 (5.2%)
Diabetes	240 (8.5%)	118 (15.1%)	22 (33.3%)	27 (25.5%)	407 (10.8%)
2012 CDC/AAP Periodontitis Classifications					
Healthy	1724 (61.4%)	266 (34.1%)	26 (39.4%)	40 (37.7%)	2056 (54.7%)
Mild	60 (2.1%)	12 (1.5%)	2 (3.0%)	0 (0%)	74 (2.0%)
Moderate	823 (29.3%)	314 (40.3%)	29 (43.9%)	53 (50.0%)	1219 (32.4%)
Severe	202 (7.2%)	187 (24.0%)	9 (13.6%)	13 (12.3%)	411 (10.9%)
2012 CDC/AAP Periodontitis Classifications (Severe or Non-severe)					
Non-severe	2607 (92.8%)	592 (76.0%)	57 (86.4%)	93 (87.7%)	3349 (89.1%)
Severe	202 (7.2%)	187 (24.0%)	9 (13.6%)	13 (12.3%)	411 (10.9%)
CVD cases					
No	2758 (98.2%)	761 (97.7%)	43 (65.2%)	77 (72.6%)	3639 (96.8%)
Yes	51 (1.8%)	18 (2.3%)	23 (34.8%)	29 (27.4%)	121 (3.2%)

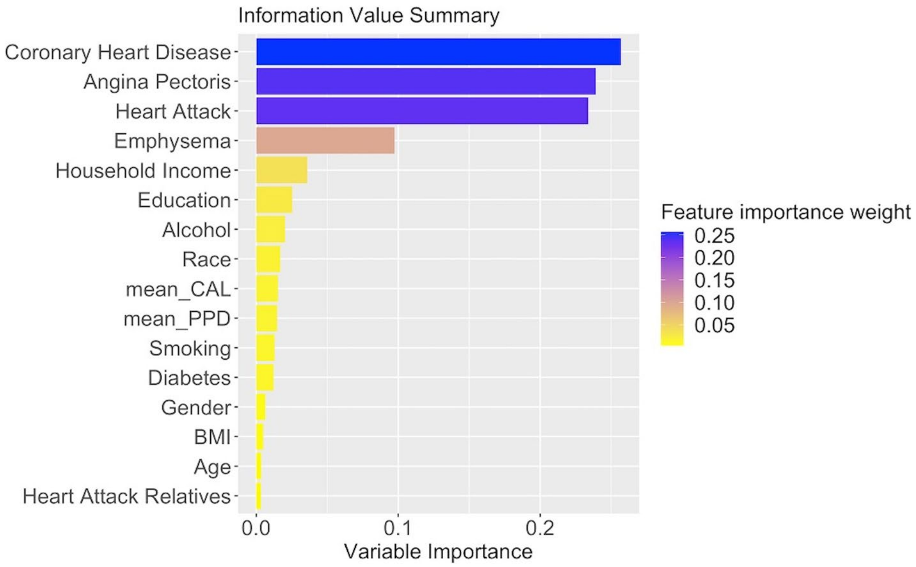

Fig. 4 Feature importance visualization to account for contributions to clustering

Table 3 The coefficient table of the regression analysis

Parameter	Parameter Estimate(B)	Std. Error	OR (95% CI)	*P*-value
Model 1	***No adjustments***			
Intercept	−3.50	0.10	/	< 2e-16 ***
PD severity	0.68	0.24	1.97 (1.55, 2.49)	0.0045**
Model 2	***Not adjust for clusters**, **only for individual characteristics***			
Intercept	0.038	0.97	/	0.97
PD severity	0.15	0.36	1.16 (0.81, 1.66)	0.68
Model 3	***Adjust for clusters***			
Intercept	−4.06	0.15	/	< 2e-16 ***
PD severity	0.67	0.27	1.95 (1.50, 2.55)	0.012*
Cluster 2	0.11	0.29	1.11 (0.84, 1.48)	0.70
Cluster 3	3.33	0.30	28.01 (20.83, 37.66)	< 2e-16 ***
Cluster 4	2.99	0.26	19.83 (15.28, 25.74)	< 2e-16 ***

relationships between different types of data. For instance, research by Elsie et al. [55] emphasized the challenges in achieving accurate clustering with mixed data types, often resulting in suboptimal interpretability. Our framework not only acknowledges these challenges but also proposes a robust solution by integrating a novel distance measurement approach that balances the influence of both data types, as well as considering feature importance to account for feature correlations. This methodological advancement directly responds to the call for more comprehensive analytical tools capable of handling the complexity of modern datasets, as discussed in the works of Amir et al. [56].

The strengths of the DAFI-Gower technique include its robust performance in scenarios with numerous redundant variables, as evidenced by simulations 4 and 5. In the empirical study using NHANES, the technique achieved superior performance as indicated by the Silhouette index. Moreover, incorporating feature importance enhances the interpretability of the clustering results, allowing better classification into high or low-risk CVD groups. Moreover, utilizing clusters in association analyses offers distinct methodological and clinical advantages. In the NHANES study, adjusting for clusters revealed significant associations, which is consistent with the well-established consensus that PD is significantly associated with CVD [47], whereas adjusting for individual confounders did not. This approach enhances the robustness and clarity of findings by reducing multicollinearity, simplifying complex relationships, and improving statistical power. In addition to investigating associations, the magnitude of the coefficients for our clusters indicates substantial variance in CVDs risk, demonstrating the potential of mixed type clustering techniques to uncover nuanced health patterns.

Despite these strengths, our study has limitations. With a time complexity of $O(n^2)$, the DAFI-Gower's runtime on NHANES datasets (3,760 observations and 18 variables) is approximately three minutes, which is longer compared to a few seconds for other methods (see Supplementary Table 9). However, it remains acceptable for most real-world clinical applications with moderate cohort sizes, and we will consider optimization strategies to improve efficiency. In addition, some methods were not included as baseline methods due to limitations of computational capacity. Two commonly used clustering methods are the Latent Class Analysis (LCA) and Latent Class Model (LCM) [57, 58], which are two statistical techniques used to identify unobserved subgroups (latent classes) within a population based on individuals' responses to observed categorical variables. Although previous research [57] showed LCM and KAMILA typically performed best in the setting of heterogeneous data, we excluded LCA and LCM in our analysis due to their largest computation time and only included KAMILA as one of the competitors.

For Gower distance, scaling Euclidean distance for continuous features can be improved using methods beyond the interquartile range (IQR). As suggested by D'Orazio [30], k-nearest neighbor [59] and kernel density estimator [60] approaches offer alternative scaling for interval and ratio scale variables. Kernel density estimator scales distances based on density, giving less weight to values in high-density regions and reducing the influence of outliers by downscaling sparse areas. K-nearest neighbor uses the average distance to each point's nearest neighbors to set a local scale, emphasizing distances within clusters and lessening the impact of extreme values. Although we do not have such variables currently, future studies could explore these methods when such data is collected, providing clinical examples.

Feature importance determination can also be refined. Currently, continuous variables are categorized by quartiles to calculate MI [19, 33], but advanced methods allow for MI estimation without discretization. For example, Brian [61] introduced a non-binning MI estimator for mixed discrete and continuous data, which can also be adapted for Jensen–Shannon divergence, providing a sophisticated measure of feature importance. Additionally, joint data reduction techniques combine dimension reduction and clustering, enabling mixed data clustering with transformed variables on comparable scales [62]. For datasets with diverse feature types, like text, approaches using distortion and convex optimization offer more complex feature weighting [63].

Future work may explore more accurate methods for calculating MI for continuous variables. In our study, algorithms were performed with default values for fairness and simplicity; however, optimizing these parameters could further enhance performance. For example, when the number of clusters is not predefined, methods such as the elbow method or prediction strength [64] can help determine an appropriate cluster count. Additionally, information-based approaches like BIC [65] could be applied in preprocessing, provided that a suitable likelihood expression is defined. Although this study includes commonly-used clustering methods, future work could expand the comparison to additional algorithms compatible with our proposed Gower distance to provide a more comprehensive evaluation. While this study is limited to hard clustering, future work could explore fuzzy clustering methods, which allow data points to belong to multiple clusters with varying degrees of membership [66], offering a more flexible clustering approach.

The practical implications of our study are significant. The DAFI-Gower technique's ability to improve cluster quality based on feature importance makes it particularly useful in clinical studies where many measures are taken, but their contributions to outcomes are unclear. Besides, using clusters in association analyses enriches our understanding of disease mechanisms and opens avenues for tailored preventive strategies addressing the multifactorial nature of chronic diseases like CVDs. This approach supports more rational patient subgroups, facilitating personalized treatment plans and improving patient outcomes. Future research should focus on refining the distance measures and feature importance calculations to further improve the technique's performance and applicability in diverse datasets.

Conclusion

This paper introduced DAFI-Gower, an innovative clustering approach for mixed-type datasets that incorporates feature importance to improve cluster quality and interpretability. DAFI-Gower provides a robust tool for identifying meaningful patterns in complex data, supporting data-driven decision-making in epidemiology and other fields requiring effective clustering.

Abbreviations

PD	Periodontitis
CVDs	Cardiovascular diseases
EHRs	Electronic health records
NHANES	National health and nutrition examination study
DAFI	Distance adjustment and feature importance
IQR	Inter-quartile range
MI	Mutual information
NMI	Normalized mutual information
PAM	Partitioning around medoids
ARI	Adjusted Rand index
KAMILA	KAymeans for MIxed Large data
CDC/AAP	Centers for Disease Control and Prevention /American Academy of Periodontology
OR	Odds ratio
CI	Confidence interval

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02427-8.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

Not applicable.

Authors' contributions
P.L. and M.A.P. conceived and designed the study. P.L. and H.Y. carried out the statistical and computational analyses. P.L. contributed to the design of the study and the interpretation of the findings. The paper was written by P.L., and revised by Y.N., H.Y., B.C., N.L., and M.A.P. All co-authors have approved the final version of the paper.

Funding
This research is supported by Duke-NUS Medical School.

Data availability
The data used in this article can be freely and openly accessed at NHANES. https://www.cdc.gov/nchs/nhanes/index.htm Accessed 9 July 2024.

Declarations

Ethics approval and consent to participate
Health information collected in the NHANES is kept in strictest confidence. During the informed consent process, survey participants are assured that data collected will be used only for stated purposes and will not be disclosed or released to others without the consent of the individual or the establishment in accordance with Sect. 308(d) of the Public Health Service Act (42 U.S.C. 242 m).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore. [2]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore. [3]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore. [4]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. [5]Institute of Data Science, National University of Singapore, Singapore, Singapore. [6]National Dental Research Institute Singapore, National Dental Centre Singapore, Singapore, Singapore.

Received: 9 July 2024 Accepted: 26 November 2024
Published online: 18 December 2024

References

1. Liu P, Wang Z, Liu N, Peres MA. A scoping review of the clinical application of machine learning in data-driven population segmentation analysis. J Am Med Inform Assoc. 2023;30(9):1573–82. https://doi.org/10.1093/jamia/ocad111.
2. Chong JL, Matchar DB. Benefits of Population Segmentation analysis for developing Health Policy to promote patient-centred care. Ann Acad Med Singap. 2017;46(7):287–9.
3. Zhou YY, Wong W, Li H. Improving care for older adults: a model to segment the senior population. Perm J. 2014 Summer;18(3):18–21.
4. Krishna K, Narasimha Murty M. Genetic K-means algorithm. IEEE Trans Syst Man Cybernetics Part B (Cybernetics). 1999;29(3):433–9.
5. Fan J, Han F, Liu H. Challenges of Big Data Analysis. Natl Sci Rev. 2014;1(2):293–314.
6. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019;7(6):442–51.
7. Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl Discov Data. 2009;3(1):1–58. https://doi.org/10.1145/1497577.1497578.
8. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. Npj Digit Med. 2023;6(1):98.
9. Wanichthanarak K, Fahrmann JF, Grapov D. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomark Insights. 2015;10(Suppl 4):1–6. https://doi.org/10.4137/BMI.S29511.
10. Shen L, Thompson PM. Brain imaging genomics: integrated analysis and machine learning. Proc IEEE. 2020;108(1):125–62.
11. Pandya S, Shah J, Joshi N, Ghayvat H, Mukhopadhyay SC, Yap MH. A novel hybrid based recommendation system based on clustering and association mining, 2016 10th International Conference on Sensing Technology (ICST). Nanjing: 2016. p. 1–6. https://doi.org/10.1109/ICSenST.2016.7796287.
12. Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. In: Machine Learning Proceedings 1995 [Internet]. Elsevier; 1995 [cited 2024 Feb 2]. pp. 194–202. https://linkinghub.elsevier.com/retrieve/pii/B9781558603776500323.
13. Ichino M, Yaguchi H. Generalized Minkowski Metrics for mixed feature-type data analysis. Syst Man Cybernetics IEEE Trans on. 1994;24:698–708.
14. MacQueen JB. Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Statistics. Berkeley: University of California Press; 1967;1:281–97. http://projecteuclid.org/euclid.bsmsp/1200512992.
15. Forgy E. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics. 1965;21:768–9.
16. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min Knowl Disc. 1998;2(3):283–304.
17. Huang JZ, Ng MK, Rong H, Li Z. Automated variable weighting in k-means type clustering. IEEE Trans Pattern Anal Mach Intell. 2005;27(5):657–68.
18. Gnanadesikan R, Kettenring JR, Tsao SL. Weighting and selection of variables for cluster analysis. J Classif. 1995;12(1):113–36.
19. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng. 2007;63(2):503–27.
20. Chae SS, Kim JM, Yang WY. Cluster analysis with Balancing Weight on mixed-type data. Commun Stat Appl Methods. 2006;13(3):719–32.
21. Lawrence CJ, Krzanowski WJ. Mixture separation for mixed-mode data. Stat Comput. 1996;6(1):85–92.
22. Browne RP, McNicholas PD. Model-based clustering, classification, and discriminant analysis of data with mixed type. J Stat Plann Inference. 2012;142(11):2976–84.
23. Hunt L, Jorgensen M. Clustering mixed data. WIREs Data Min Knowl Discov. 2011;1(4):352–61.
24. McNicholas PD, Murphy TB. Parsimonious gaussian mixture models. Stat Comput. 2008;18(3):285–96.
25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from Incomplete Data via the EM Algorithm. J Royal Stat Soc Ser B (Methodological). 1977;39(1):1–38.
26. Foss A, Markatou M, Ray B, Heching A. A semiparametric method for clustering mixed data. Mach Learn. 2016;105(3):419–58.
27. Chu Ctao, Kim S, Lin Y, an, Yu Y, Bradski G, Olukotun K et al. Map-Reduce for Machine Learning on Multicore. In: Advances in Neural Information Processing Systems. MIT Press; 2006 [cited 2024 Jan 29]. https://papers.nips.cc/paper_files/paper/2006/hash/77ee3bc58ce560b86c2b59363281e914-Abstract.html.
28. Wolfe J, Haghighi A, Klein D. Fully distributed EM for very large datasets. In: Proceedings of the 25th international conference on Machine learning. New York, NY, USA: Association for Computing Machinery; 2008 [cited 2024 Jan 29]. pp. 1184–91. (ICML '08). Available from: https://doi.org/10.1145/1390156.1390305.
29. Gower JC. A General Coefficient of Similarity and some of its Properties. Biometrics. 1971;27(4):857–71.
30. D'Orazio M. Distances with mixed type variables some modified Gower's coefficients. ArXiv. 2021 Jan 7 [cited 2024 Jan 2]; https://www.semanticscholar.org/paper/Distances-with-mixed-type-variables-some-modified-D%27Orazio/e702062429d9642bc12ac5f79bd71645aeaa8dd0.
31. Pinto A, Faiz O, Davis R, Almoudaris A, Vincent C. Surgical complications and their impact on patients' psychosocial well-being: a systematic review and meta-analysis. BMJ Open. 2016;6(2):e007224.
32. Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Applic. 2014;24(1):175–86.

33. Ji J, Pang W, Zhou C, Han X, Wang Z. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowl Based Syst. 2012;30:129–35.
34. Singh V, Verma NK. An Entropy-based Variable Feature Weighted Fuzzy k-Means Algorithm for High Dimensional Data. arXiv; 2019 [cited 2024 Jan 4]. Available from: http://arxiv.org/abs/1912.11209.
35. Brown G, Pocock A, Zhao MJ, Luján M. Conditional likelihood maximisation: a Unifying Framework for Information Theoretic feature selection. J Mach Learn Res. 2012;13(2):27–66.
36. Yin L, Xingfei M, Mengxi Y, Wei Z, Wenqiang G. Improved Feature Selection Based on Normalized Mutual Information. In: 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). 2015 [cited 2024 Jan 15]. pp. 518–22. Available from: https://ieeexplore.ieee.org/document/7429669.
37. Horibe Y. Entropy and correlation. IEEE Trans Syst Man Cybernetics. 1985;SMC–15(5):641–2.
38. Chen S, Ma B, Zhang K. On the similarity metric and the distance metric. Theor Comput Sci. 2009;410(24):2365–76.
39. Mousavi E, Sehhati M. A generalized multi-aspect distance metric for mixed-type data clustering. Pattern Recogn. 2023;138:109353.
40. Botyarov M, Miller EE. Partitioning around medoids as a systematic approach to generative design solution space reduction. Results Eng. 2022;15:100544.
41. Van der Laan M, Pollard K, Bryan J. A new partitioning around medoids algorithm. J Stat Comput Simul. 2003;73(8):575–84.
42. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218.
43. Szepannek G, clustMixType. User-friendly clustering of mixed-type data in R. R J. 2019;10(2):200.
44. Foss AH, Markatou M. Kamila: clustering mixed-type data in R and Hadoop. J Stat Softw. 2018;83:1–44.
45. Chaudhuri A, Samanta D, Sarma M. Two-stage approach to feature set optimization for unsupervised dataset with heterogeneous attributes. Expert Syst Appl. 2021;172:114563.
46. DAFI-Gower Source Codes in Github. https://github.com/Pinyan-Liu/DAFI-Gower-Distance
47. Shetty B, Fazal I, Khan SF, Nambiar M, Prasad DKI. Association between cardiovascular diseases and periodontal disease: more than what meets the eye. Drug Target Insights. 2023;17:31–8.
48. Eke PI, Page RC, Wei L, Thornton-Evans G, Genco RJ. Update of the Case definitions for Population-based surveillance of Periodontitis. J Periodontol. 2012;83(12):1449–54.
49. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. 2009. Available from: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2009.
50. Centers for disease control and prevention, national center for health statistics. National health and nutrition examination survey. Available: www.cdc.gov/nchs/nhanes/about_nhanes. html [Accessed 1 Feb 2024].
51. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806–8.
52. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of Coronary Heart Disease in Low- and Middle-Income Countries. Curr Probl Cardiol. 2010;35(2):72–115.
53. Rousseeuw PJ, Silhouettes. A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
54. Batool F, Hennig C. Clustering with the average Silhouette Width. Comput Stat Data Anal. 2021;158:107190.
55. Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of Clustering Multimodal Clinical Data: review of applications in Asthma Subtyping. JMIR Med Inf. 2020;8(5):e16452.
56. Ahmad A, Khan SS. Survey of State-of-the-art mixed data clustering algorithms. IEEE Access. 2019;7:31883–902.
57. Marbac M, Sedki M. VarSelLCM: an R/C++package for variable selection in model-based clustering of mixed-data with missing values. Wren J, editor. Bioinformatics. 2019;35(7):1255–7.
58. Vermunt JK, Magidson J. Latent class cluster analysis. In Hagenaars J, McCutcheon A. (Eds.), Applied latent class analysis. Cambridge University Press; 2002. p. 89–106.
59. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69(6):066138.
60. Moon YI, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top. 1995;52(3):2318–21.
61. Ross BC. Mutual Information between Discrete and Continuous Data Sets. Marinazzo D, editor. PLoS ONE. 2014;9(2):e87357.
62. van de Velden M, Iodice D'Enza A, Markos A. Distance-based clustering of mixed data. WIRE Comput Stat. 2019;11(3):e1456.
63. Modha DS, Spangler WS. Feature weighting in k-Means clustering. Mach Learn. 2003;52(3):217–37.
64. Tibshirani R, Walther G. Cluster validation by Prediction Strength. J Comput Graphical Stat. 2005;14(3):511–28.
65. Schwarz G. Estimating the dimension of a model. Annals Stat. 1978;6(2):461–4.
66. Tortora C, Palumbo F. Clustering mixed-type data using a probabilistic distance algorithm. Appl Soft Comput. 2022;130:109704.

Publisher's Note