POLITECNICO DI TORINO
Repository ISTITUZIONALE

Comparison of different similarity measures in hierarchical clustering

Original
Comparison of different similarity measures in hierarchical clustering / Vagni, Marica; Giordano, Noemi; Balestra, Gabriella; Rosati, Samanta. - ELETTRONICO. - (2021), pp. 1-6. (Intervento presentato al convegno 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA) tenutosi a Lausanne, Switzerland nel 23-25 June 2021) [10.1109/MeMeA52024.2021.9478746].

Availability:
This version is available at: 11583/2913759 since: 2021-09-15T16:49:25Z

Publisher:
IEEE

Published
DOI:10.1109/MeMeA52024.2021.9478746

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

12 October 2024

Comparison of different similarity measures in hierarchical clustering

Marica Vagni
Department of Electronics and Telecommunications
Politecnico di Torino
Torino, Italy
marica.vagni@polito.it

Noemi Giordano
Department of Electronics and Telecommunications
Politecnico di Torino
Torino, Italy
noemi.giordano@polito.it

Gabriella Balestra
Department of Electronics and Telecommunications
Politecnico di Torino
Torino, Italy
gabriella.balestra@polito.it

Samanta Rosati
Department of Electronics and Telecommunications
Politecnico di Torino
Torino, Italy
samanta.rosati@polito.it

***Abstract*—The management of datasets containing heterogeneous types of data is a crucial point in the context of precision medicine, where genetic, environmental, and life-style information of each individual has to be analyzed simultaneously. Clustering represents a powerful method, used in data mining, for extracting new useful knowledge from unlabeled datasets. Clustering methods are essentially distance-based, since they measure the similarity (or the distance) between two elements or one element and the cluster centroid. However, the selection of the distance metric is not a trivial task: it could influence the clustering results and, thus, the extracted information. In this study we analyze the impact of four similarity measures (Manhattan or L1 distance, Euclidean or L2 distance, Chebyshev or L∞ distance and Gower distance) on the clustering results obtained for datasets containing different types of variables. We applied hierarchical clustering combined with an automatic cut point selection method to six datasets publicly available on the UCI Repository. Four different clusterizations were obtained for every dataset (one for each distance) and were analyzed in terms of number of clusters, number of elements in each cluster, and cluster centroids. Our results showed that changing the distance metric produces substantial modifications in the obtained clusters. This behavior is particularly evident for datasets containing heterogeneous variables. Thus, the choice of the distance measure should not be done a-priori but evaluated according to the set of data to be analyzed and the task to be accomplished.**

Keywords—similarity measures, distance measures, Gower distance, hierarchical clustering, UCI repository

I. Introduction

Precision Medicine (PM) aims to develop treatments and preventions strategies based on genetic, environmental and life-style information of each specific individual [1]. To this scope, large amount of data has to be collected and analyzed in order to retrieve valuable information. These datasets often contain heterogeneous types of variables such as binary, continuous, categorical and integer variables, and parameters calculated from signals and images.

Data mining (DM) is defined as "the extraction of meaningful knowledge from useful but non-evident information which is hidden within large datasets" [2]. Focusing on the clinical context, DM techniques are mainly applied with three objectives: understanding the clinical data, assisting healthcare professionals, and developing a data analysis methodology suitable for medical data [3]. Moreover, in the era of PM, these techniques are supposed to be able to process heterogeneous sets of data in an efficient manner [4].

Unsupervised learning represents the class of algorithms mainly used for DM, allowing for extracting new knowledge from unlabeled datasets. In particular, clustering aims to divide the dataset into groups (*clusters*) consisting of similar elements. Similarity is demonstrated by low intra-cluster variability and high inter-cluster distance. Even if clustering is mainly used for clinical dataset analysis, some applications for signals and image analysis have been proposed. In a previous study, a hierarchical clustering technique was successfully applied to EMG signals to obtain information about the muscle activations that are necessary for the biomechanical function of walking [5], [6]. Rosati et al. [7] proved that clustering applied to medical images can improve cancer detection.

Clustering techniques can be mainly divided into two groups: partitional algorithms and hierarchical algorithms. The first group of algorithms (such as *kmeans*) divides data into a set of *k* disjoint clusters, whereas the second group aims to build a hierarchy of nested clusters through a bottom-up (agglomerative) or a top-down (divisive) approach. Both partitional and hierarchical algorithms are based on the assessment of the similarity between two elements in the dataset or one element and the cluster centroid. Similarity is usually measured by means of a distance metric: small distance corresponds to high similarity and vice versa.

The selection of the distance metric is not a trivial task, both for clustering and for distance-based classification. Several measurements have been proposed in literature [8], unfortunately the majority of them are suitable only in case of continuous or integer variables. In a recent review, the authors analyzed the effect of different similarity measures on the classification performances of a KNN classifier [9]. A similar analysis was conducted by Dos Santos and Zarate [10], that compared different measures for clustering categorical data.

Focusing on datasets with heterogeneous types of variables, the problem of selecting the appropriate similarity measure is challenging for two reasons: (a) the different meaning that a specific value assumes for different types of variables (for example the number "1" for continuous or binary variables); (b) how to manage together numerical (ordinal) variables and categorical (ordinal or not) variables. Some studies faced this last problem simply transforming continuous variables into discrete ones [10], [11]. In [12] the problem of heterogeneous datasets was dealt with, comparing three different measures for computing Mahalanobis-type distances for classification and principal components analysis applied to categorical and mixed variables. To the best of our

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

knowledge, no similar studies have been conducted about clustering.

The aim of this study is to analyze the impact of different similarity measures on the clustering results obtained for datasets containing different types of variables. To this scope, we used six datasets publicly available on the UCI Repository (https://archive.ics.uci.edu/ml/index.php) and we compared the results of hierarchical clustering using four similarity metrics

II. Materials and Methods

A. Datasets Description

The six datasets publicly available on the UCI Repository [13] were: Breast Cancer [14], Mammographic Mass [15], Breast Cancer Wisconsin (Diagnostic), SPECT Heart, HCV Data, Heart Disease (Cleveland) [16]. All of them belong to the Life Science area.

They were selected in order to include in our analysis different types of variables (real, integer, categorical). Both homogeneous and heterogeneous datasets were chosen, to better understand the characteristics of the different similarity measures. Moreover, to obtain a robust analysis of our results, a further selection criterion was set on the number of elements that must be higher than 200.

From the six selected datasets, we decided to completely remove those elements containing at least one missing value (MV), because the management of MVs is out of the purpose of this study. Moreover, even if several study demonstrated that a proper selection of the features can improve the system performances [17], [18], in this study we decided to consider the entire set of features to avoid bias due to the feature selection method. Finally, since the great majority of datasets in the Life Science area of UCI repository are supervised (including the 6 selected here), the information related to the class was ignored during the clustering analysis.

The characteristics of the datasets, after MV removal, are summarized in Table I.

B. Similarity (or Distance) Measures

In this study four similarity measures were used to compare two elements x and y characterized by m variables.

- *Cityblock* or *Manhattan* or *L1 distance.* It is the sum of the absolute differences between x and y on each variable obtained as:

$$dist_{L1(x,y)} = \sum_{i=1}^{m} |x_i - y_i| \quad (1)$$

- *Euclidean* or *L2 distance.* It is the most common used distance measure, and it calculates the square root of sum of the squared differences between x and y on each variable:

$$dist_{L2(x,y)} = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \quad (2)$$

- *Chebyshev* or *L∞ distance.* It corresponds to the maximum absolute difference between x and y on all variables:

$$dist_{L\infty(x,y)} = \max_{m}(|x_i - y_i|) \quad (3)$$

- *Gower distance* [19]. It is the best-known dissimilarity measure for mixed data. For two elements x and y, the distance is expressed as:

$$dist_{Gower(x,y)} = \sum_{i=1}^{m} d_i(x_i, y_i) \quad (4)$$

where

$$d_i(x_i, y_i) = \begin{cases} 1 & if\ x_i \neq y_i \\ 0 & if\ x_i = y_i \end{cases} \quad (5)$$

if the i-th variable is categorical, or:

$$d_i(x_i, y_i) = \frac{|x_i - y_i|}{R_i} \quad (6)$$

if the i-th variable is quantitative (real or integer), where R_i represents the range of the variable.

Since the variables included in each dataset had very different ranges and this can influence the distance measures (variables having larger ranges will dominate over the others), each dataset was normalized before applying clustering. We used the min-max scaling normalization to obtain all variables between 0 and 1. For each variable i, the normalized value was obtained as:

$$Var_norm_i = \frac{Var_i - \min(Var_i)}{\max(Var_i) - \min(Var_i)}$$

where Var_i is the original value of the i-th variable. We preferred to use the min-max scaling instead of *z-score* (or *standard score*) normalization, to preserve the original value distribution of each variable.

C. Clustering

Agglomerative hierarchical clustering was applied to each dataset. Starting with each element considered as a cluster, agglomerative hierarchical clustering iteratively merges the two most similar clusters, until all elements are pulled together in a single cluster. This iterative process is commonly depicted as a tree (called *dendrogram*) and final clusters are identified by cutting the tree at a certain level.

In this study we use the *complete linkage* as method for selecting the two clusters to be merged at each iteration. This means that, in each iteration, the farthest distance between every pair of elements in two clusters is considered as inter-

TABLE I. Characteristics of the Six Datasets

Dataset Name	# of elements	# of Attributes			
		Real	Integer	Catego-rical	Total
Breast Cancer	277	0	0	9	9
Mammographic Mass	860	0	1	3	4
Breast Cancer Wisconsin (Diagnostic)	569	30	0	0	30
SPECT Heart	267	0	44	0	44
HCV Data	589	10	1	1	12
Heart Disease (Cleveland)	297	1	4	8	13

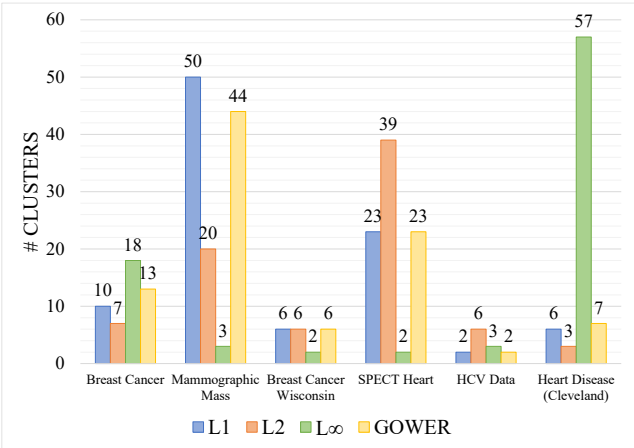

Fig. 1. Number of clusters identified with the different distances on the six datasets

cluster distance, and the two clusters with the smallest farthest distance are joined together.

The four distance measures were used to construct four dendrograms for each dataset. For each dendrogram, the final clusters were identified using the automatic method for cut point identification proposed in the application presented in [6]. It aims to reduce the intra-cluster variability by comparing three different cut points and selecting the best one.

We decided to use hierarchical clustering combined with the automatic cut point identification in order to avoid any subjective selection of clustering parameters, such as number of clusters and cut point, that could influence the final results.

D. *Results Analysis*

The results of the hierarchical clustering using the four similarity measures were compared, for each dataset, in terms of:

- Number of clusters,
- Number of elements in each cluster,
- Centroid of each cluster, calculated as median value of the elements within the cluster, for each variable.

In particular, to simplify the visualization and the results analysis, we distinguished between significant clusters (i.e. containing at least 10% of the total number of elements) and not significant clusters.

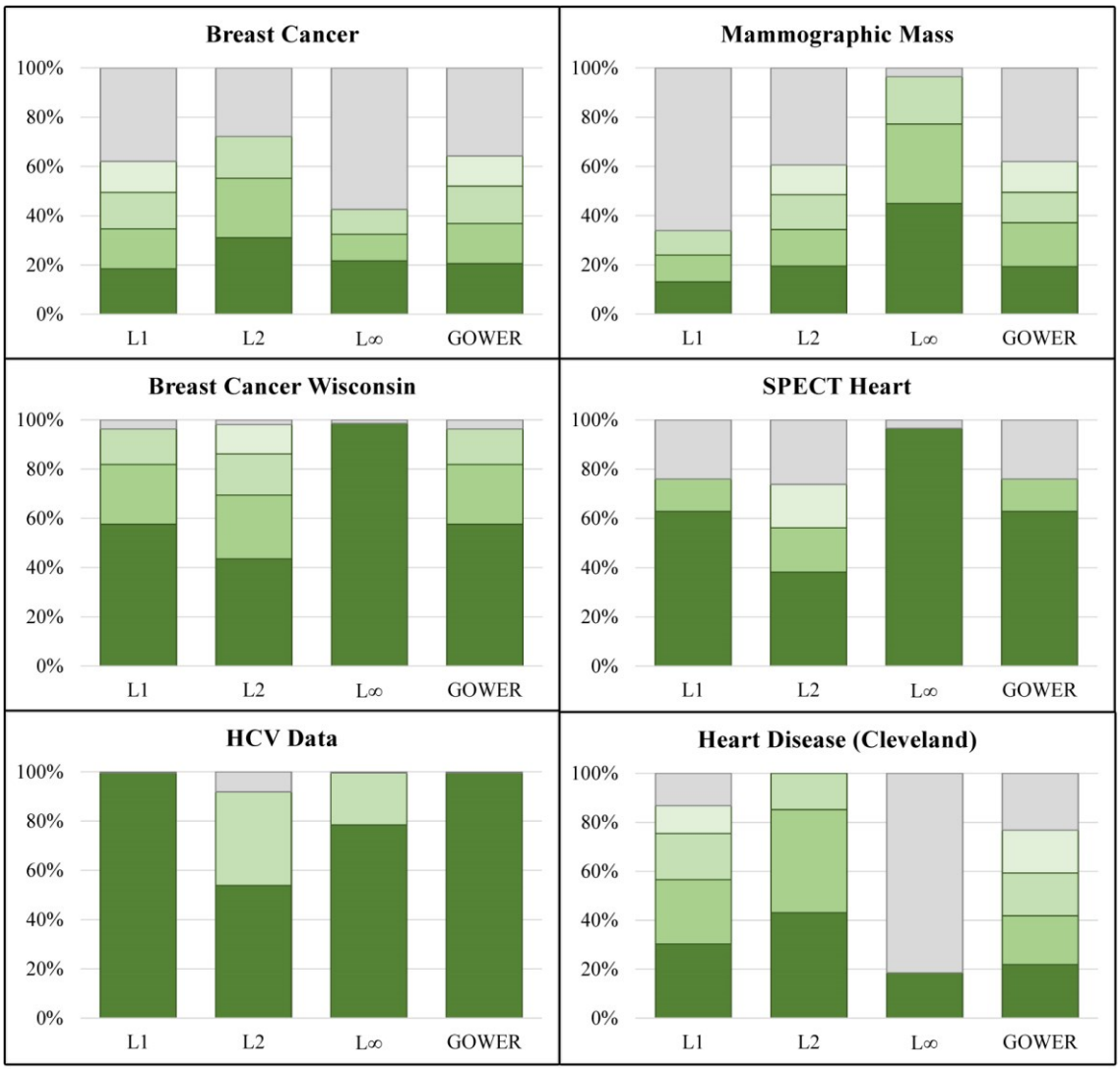

Fig. 2. Percentage of elements contained in significant clusters (in green) and in not-significant clusters (in grey), for each dataset and for the four distance measures.

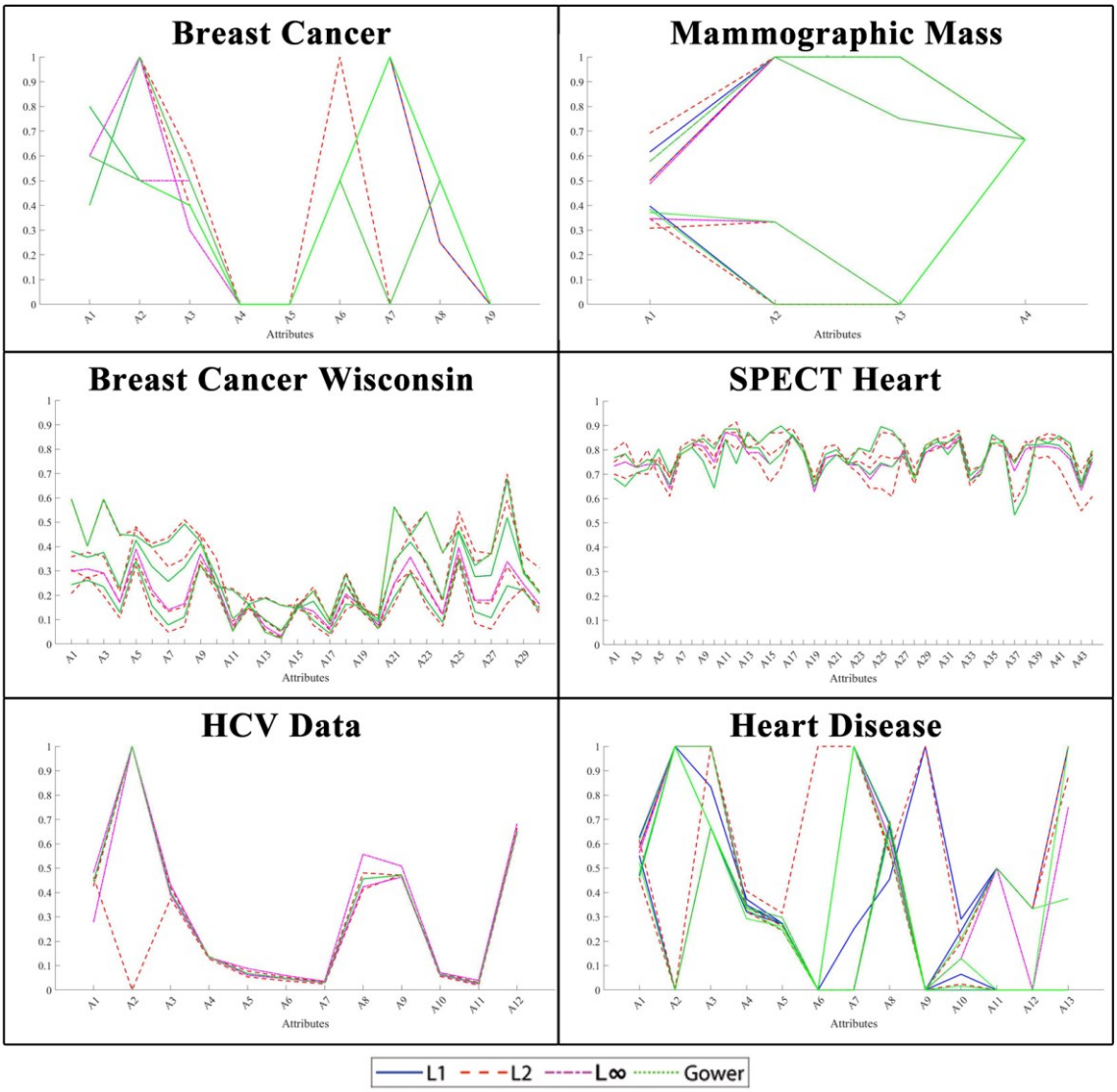

Fig. 3. Centroids of significant clusters, for each dataset and distance measure.

III. Results and Discussion

In this study we analyze the effect of four distance measures on the clustering results obtained for six datasets containing different types of variables. Three measures (L1, L2 and L∞) are usually used for numerical (real or integer) variables and homogeneous datasets, whereas Gower distance was previously found as the best measure for managing categorical variables [10].

Fig. 1 shows the number of clusters identified for each dataset, using the four distance measures. As it emerges from the figure, changing the distance measure could significantly modify the number of identified clusters. In particular, lower variability in the number of clusters can be observed for Breast Cancer Wisconsin and HCV Data. It must be noticed that the first dataset contains only continuous numerical variables, while the second includes 11 numerical variables (10 real and 1 integer) and 1 categorical variable.

Fig. 2 presents the percentage of elements contained in significant clusters (in green) and in not-significant clusters (in grey), for each couple dataset-distance. From this analysis it emerges that L1 and Gower distances allow to obtain a similar number of elements in significant clusters, and also a similar number of clusters (see Fig. 1). This can be explained looking at the definition of the Gower distance showed in (4), that corresponds to L1 distance in (1) in case of quantitative (real or integer) attributes normalized between 0 and 1 (as in this case). Conversely, the L∞ is the distance that mostly deviates from the others from the point of view of the number of clusters and their numerousness.

Fig. 3 shows the centroids of the significant clusters obtained with different similarity measures (represented with different line colors) applied to each dataset. Even if there are situations in which the values of the centroids are very similar using different distances (e.g. HCV Data), in most cases the centroids of significant clusters assume different values by changing the distance measure, meaning that these clusters include very different elements. This behavior is particularly evident for Heart Disease dataset, which is the most heterogeneous dataset among the six included in this study. In fact, analyzing the dendrograms obtained for this dataset (Fig. 4) with different distance measures, the shape of the tree changes in a significant way. This means that the hierarchical clustering finds a couple of most similar clusters to be merged at each iteration that is completely different changing the

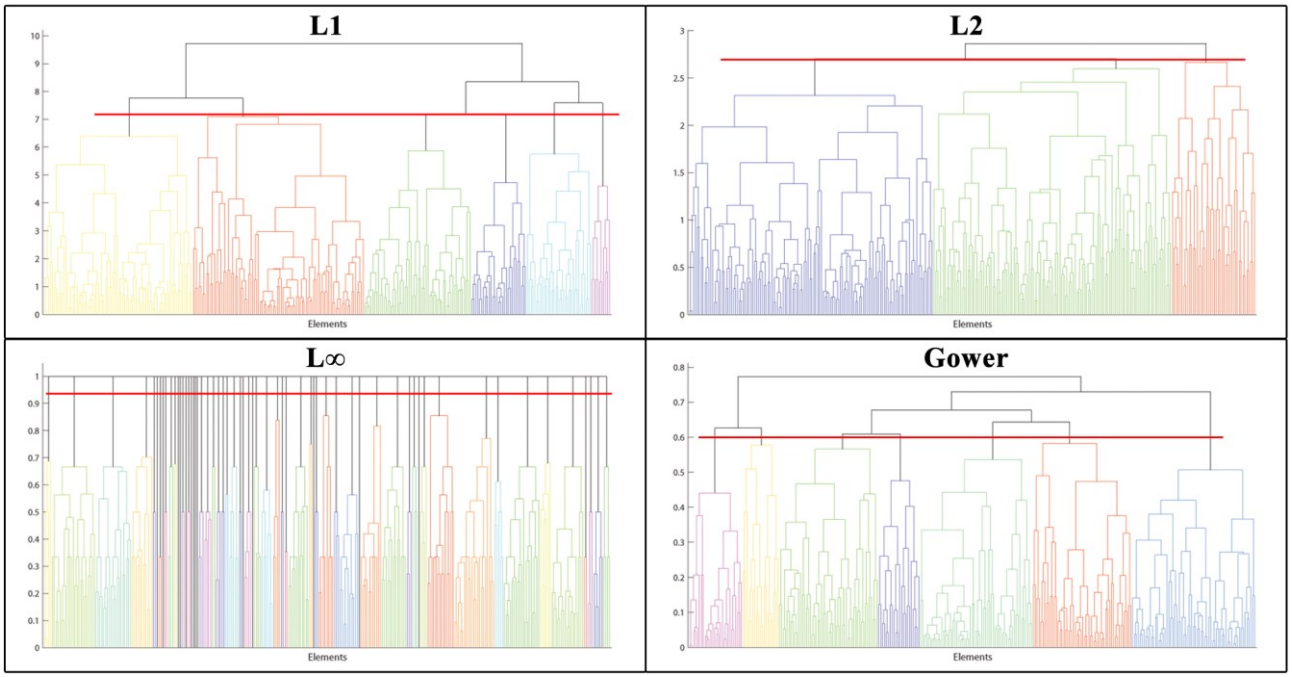

Fig. 4. Dendrograms obtained for Heart Disease dataset using the four distance measures. The red horizontal lines show the cut points automatically identified for each tree and the resulting clusters are highlighed with different colors.

distance measure. A similar behavior can be observed also for the other datasets.

All these findings reveal that the choice of the distance measure to be used for clustering is not a trivial task, since each metric returns very different groups of elements. This is particularly evident when dealing with heterogeneous sets of variables, as in Heart Disease dataset. Moreover, even if the L2 or Euclidean distance is the expected choice in most studies, it could not be always the right choice.

IV. Conclusions

In this study we analyzed the impact of four different similarity measures in dataset with heterogeneous types of variables on hierarchical clustering results. We used six datasets publicly available and containing different types of variables. We applied hierarchical clustering to each dataset using four different similarity measures that are commonly used. We compared the results in terms of number of clusters, number of elements in each cluster and cluster centroids.

From our findings it emerged that changing the distance metric produces substantial modifications in the obtained clusters. This is particularly evident for datasets containing heterogeneous types of variables. From the DM point of view, this means that different information and knowledge will be extracted.

Thus, we can conclude that the choice of the distance measure should not be done a-priori but evaluated according to the set of data to be analyzed and the task to be accomplished.

References

[1] Z. G. Wang, L. Zhang, and W. J. Zhao, “Definition and application of precision medicine,” Chinese Journal of Traumatology - English Edition. 2016.

[2] A. M. Jimenez-Carvelo, “Data mining/machine learning methods in foodomics,” Current Opinion in Food Science. 2021.

[3] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, “Clinical data mining: a review.,” Yearb. Med. Inform., pp. 121–33, 2009.

[4] S. J. Lee and K. Siau, “A review of data mining techniques,” Industrial Management and Data Systems. 2001.

[5] S. Rosati, V. Agostini, M. Knaflitz, and G. Balestra, “Muscle activation patterns during gait: A hierarchical clustering analysis,” Biomed. Signal Process. Control, vol. 31, pp. 463–469, 2017.

[6] S. Rosati, C. Castagneri, V. Agostini, M. Knaflitz, and G. Balestra, “Muscle contractions in cyclic movements: Optimization of CIMAP algorithm,” in Proceedings of the Annual International Conference of the IEEE Engineering in *Medicine and Biology Society, EMBS*, 2017.

[7] S. Rosati, V. Giannini, C. Castagneri, D. Regge, and G. Balestra, “Dataset homogeneity assessment for a prostate cancer CAD system,” in 2016 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016 - Proceedings, 2016.

[8] D. J. Weller-Fahy, B. J. Borghetti, and A. A. Sodemann, “A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection,” IEEE Communications Surveys and Tutorials. 2015.

[9] H. A. Abu Alfeilat et al., “Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review,” Big Data. 2019.

[10] T. R. L. Dos Santos and L. E. Zárate, “Categorical data clustering: What similarity measure to recommend?,” Expert Syst. Appl., 2015.

[11] S. Boriah, V. Chandola, and V. Kumar, “Similarity measures for categorical data: A comparative evaluation,” in Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics 130, 2008.

[12] B. McCane and M. Albert, “Distance functions for categorical and mixed variables,” Pattern Recognit. Lett., 2008.

[13] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: http://archive.ics.uci.edu/ml.

[14] M. Zwitter and M. Soklic, “UCI Machine Learning Repository: Breast Cancer Data Set,” https://archive.ics.uci.edu/ml/datasets/Breast+Cancer. 2009.

[15] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, “The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process,” Med. Phys., 2007.

[16] R. Detrano, “UCI Machine Learning Repository: Heart Disease Data Set,” https://archive.ics.uci.edu/ml/datasets/Heart+Disease, 2019. .

[17] S. Rosati, G. Balestra, and F. Molinari, “Feature selection applied to the time-frequency representation of muscle near-infrared spectroscopy (NIRS) signals: Characterization of diabetic oxygenation patterns,” J. Mech. Med. Biol., vol. 12, no. 4, 2012.

[18] S. Rosati, G. Balestra, and F. Molinari, “Feature Extraction by QuickReduct Algorithm: Assessment of Migraineurs Neurovascular Pattern,” J. Med. Imaging Heal. Informatics, vol. 1, no. 2, pp. 184–192, 2011.

[19] J. C. Gower, “A General Coefficient of Similarity and Some of Its Properties,” Biometrics, 1971.