Multi-view Clustering of Heterogeneous Health Data: Application to Systemic Sclerosis

DOI:
10.1007/978-3-031-14721-0_25

Document Version
Accepted author manuscript

Published in:
Parallel Problem Solving from Nature – PPSN XVII

Multi-view clustering of heterogeneous health data: Application to systemic sclerosis

Adán José-García[1], Julie Jacques[1,2], Alexandre Filiot[3], Julia Handl[6], David Launay[4], Vincent Sobanski[3,5], Clarisse Dhaenens[1]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[2] FGES, Université Catholique de Lille, F-59000 Lille, France
[3] Univ. Lille, Inserm, CHU Lille, U1286, INFINITE, F-59000 Lille, France
[4] Univ. Lille, Inserm, CHU Lille, Service de Médecine Interne et Immunologie Clinique, CeRAINO, U1286, INFINITE, F-59000 Lille, France
[5] Institut Universitaire de France (IUF)
[6] Alliance Manchester Business School, University of Manchester, Manchester, UK

Abstract. Electronic health records (EHRs) involve heterogeneous data types such as binary, numeric and categorical attributes. As traditional clustering approaches require the definition of a single proximity measure, different data types are typically transformed into a common format or amalgamated through a single distance function. Unfortunately, this early transformation step largely pre-determines the cluster analysis results and can cause information loss, as the relative importance of different attributes is not considered. This exploratory work aims to avoid this premature integration of attribute types prior to cluster analysis through a multi-objective evolutionary algorithm called `MVMC`. This approach allows multiple data types to be integrated into the clustering process, explore trade-offs between them, and determine consensus clusters that are supported across these data views. We evaluate our approach in a case study focusing on systemic sclerosis (SSc), a highly heterogeneous auto-immune disease that can be considered a representative example of an EHRs data problem. Our results highlight the potential benefits of multi-view learning in an EHR context. Furthermore, this comprehensive classification integrating multiple and various data sources will help to understand better disease complications and treatment goals.

Keywords: Clustering · Multi-view clustering · Systemic sclerosis · Multi-objective optimization.

1 Introduction

Many real-world applications consist of heterogeneous datasets comprising multiple attribute types, including binary, numerical, and categorical features. For example, electronic health records (EHRs) in medicine consist of heterogeneous structured and unstructured data elements, including demographic information, diagnoses, laboratory results, medication prescriptions, and free-text clinical

notes [37,28]. In this regard, unsupervised machine learning methods are often used to discover homogeneous groups from unlabeled data because limited information is known about the classes' distribution in these heterogeneous datasets. However, most clustering algorithms are limited to working on a single specific attribute type (i.e. numerical or nominal).

Two approaches are mainly used to address this heterogeneous data clustering problem: (i) methods based on features transformation such as discretization and (ii) methods that directly use a proximity measure designed to handle mixed-attribute types such as the Gower distance. Despite their popularity, those approaches either yield substantial information loss (i) or require the selection of the "best" proximity measure beforehand (ii).

This work explores multi-view clustering to integrate multiple attribute types (data views) into the clustering process. First, specialized dissimilarity measures are used to create views, each characterized by a specific attribute type in the heterogeneous dataset. Then, the multi-view clustering algorithm explores trade-offs between the views to discover consensus clusters supported across all views. This approach was applied and evaluated in a case study of systemic sclerosis (SSc), a highly heterogeneous disease that can be considered a representative example of an EHRs data problem.

2 Background and Related Work

With the advent of so-called *big-data*, most real-world problems now involve multiple, heterogeneous data sources. Dealing with mixed types of attributes remains challenging for the clustering and clinical communities as conventional clustering algorithms require a single common data format (e.g. numerical or categorical). In the present section, we look at this heterogeneous data clustering problem through the lens of distance-based methods. A more complete, exhaustive review of other research fields, e.g., hierarchical ([19,13]), model-based ([22,6,16]) and neural network-based clusterings ([5]), will be addressed in future work. With this in mind, we recall that no single *best* clustering method exists in a general sense [15,17,36], but rather a wide variety of clustering techniques that must be carefully selected depending on the data at hand, especially in a clinical setting.

2.1 Distance-Based Clustering on Heterogeneous Data

Most conventional, e.g., distance-based clustering algorithms work with numerical-only or categorical-only data. Two main approaches are usually followed to deal with mixed-type data [38,15,40,3]: (i) methods based on features transformation [7,41,8] and (ii) methods that cluster the heterogeneous data types directly [21,29,9,18,2,10,39].

Data transformation-based methods aim to first unify the data format and then apply a distance-based clustering method, such as K-means [38]. It consists in either discretizing numerical variables into nominal ones (needed for K-modes

clustering) or reciprocally encoding nominal attributes into continuous ones (needed for K-means clustering). Although those transformations are commonly used for clustering, it involves a potentially substantial information loss, as the clustering results strongly rely on either the cut-points (which may be inappropriate) or the coding mechanism and its underlying assumptions. Alternative approaches have been proposed to address this limitation. Wei et al. [41] proposed a mutual information-based unsupervised feature transformation (UFT) for non-numerical variables, avoiding the need for manual coding. Another popular approach is to use dimensionality reduction techniques, such as Factor Analysis of Mixed Data [8], in complement to some clustering techniques.

On the other hand, most distance-based clustering methods use a single proximity measure designed to handle mixed-data types [21,29,9,18,2,10]. The Gower distance is a widespread example of such a measure, which may be best suited depending on the data clustering structure [9]. Ahmad et al. [2] proposed a K-means algorithm based on a weighted combination of the Euclidean distance and the co-occurrence of discrete values, addressing some limitations of previous K-prototypes algorithm from Huang et al. [21]. Further work has been published by Ahmad et al. [4] on a novel K-means initialization technique for mixed data, called *initKmix*, which may outperform random initialization methods on several heterogeneous datasets. Recently, Budiaji et al. [10] proposed a simple and fast K-medoids algorithm (SFKM) combined with a generalized distance function (GDF), allowing more flexible trade-offs between numerical, binary, and categorical variables. Similarly, Harikumar and Surya [18] proposed a K-medoids approach based on a similarity measure in the form of a triplet. Among the wide range of mixed-types-based proximity measures, one can also cite the work of Li et al. [29] focusing on similarity-based agglomerative clustering (SBAC), an algorithm based on the Goodall dissimilarity.

For a given dataset, most of the above methods require the selection of the "best" proximity measure (or "best" weighting of distinct proximity measures) in advance. Therefore, finding more generic, adaptive trade-offs between the contributions of the different data types remains challenging. Multi-view clustering [27,1] potentially addresses these limitations by dividing the dataset into subsets, called *views*, each characterized by a given data type, and then treats them simultaneously. In this work, we explore the use of multi-view clustering to integrate multiple data views during the clustering process.

2.2 From Single to Multi-Objective Clustering

In view of the complementarity between different distance functions, the optimal cluster structures could be better identified using multiple proximity measures *simultaneously* [31,30,11,14,26,27,25]. As said, traditional clustering algorithms require the choice of a single proximity measure such as the Euclidean, Hamming or Cosine distance. One approach is to assign weights to the different proximity measures [21,6,20,12]. However, the appropriate weighting is hard to determine without any prior knowledge of the data itself, and the reliability of the information provided by the distance measures.

Multi-view clustering algorithms can integrate multiple dissimilarity matrices simultaneously in order to find consensus clusters that are consistent across the different data views [27,14], and yield high-quality clustering results that optimally balance the contribution of each data source [26]. Recent research has reported some first steps to exploit the intrinsic multi-criterion nature of the multi-view problems [31,30,26,27,25].

Liu et al. [31] presented a multi-objective evolutionary algorithm (based on NSGA-II [30]) that simultaneously considers two different distance measures (Euclidean and Path distances). Each individual is represented using a label-based encoding of size N (number of data points) which is then evaluated using the intra-cluster variance with respect to both distance measures. Afterward, Liu et al. [30] extended this work by proposing a fuzzy clustering approach based on a multi-objective differential evolution algorithm. In this approach, a centroid-based codification is used to represent the candidate clustering solutions. However, these methods are currently limited to two views due to the lack of generality of the Pareto dominance-based approaches. In this regard, Jose-Garcia et al. [27,25] proposed a many-objective approach to multi-view data clustering that exploits the benefits of complementary information sources taken from multiple dissimilarity matrices. Additionally, this multi-view clustering algorithm allows scaling with respect to the number of data views.

3 Multi-view Clustering Approach

The proposed methodology aims to provide a solution in the context of cluster analysis to deal with heterogeneous data characterized by multiple attribute types. First, the data is decomposed into several subsets according to the attribute types. Subsequently, a suitable proximity measure is chosen for each data subset generating a dissimilarity matrix. Finally, a multi-objective evolutionary clustering algorithm uses all dissimilarity matrices as data views to find consensus clusters across the data views. This approach is illustrated in a general way in Figure 1 and described in detail in the following sections.

3.1 Construction of the Data Views

Multi-view clustering algorithms use multiple feature spaces (data views) simultaneously. The construction and selection of data views is an important step for the accurate functioning of the algorithm. In this setting, each view represents a given data source that describes a specific perspective of a phenomenon. In this regard, in the presence of a heterogeneous dataset, we propose to create different views for different types of attributes, e.g. binary, numerical and categorical. Therefore, the database is decomposed into subsets of attributes according to their data types, resulting in many feature spaces. Then, for each data-type feature space, an appropriate proximity measure is used to generate a dissimilarity matrix representing a particular data view of the overall heterogeneous problem. To the best of our knowledge, this is the first time an

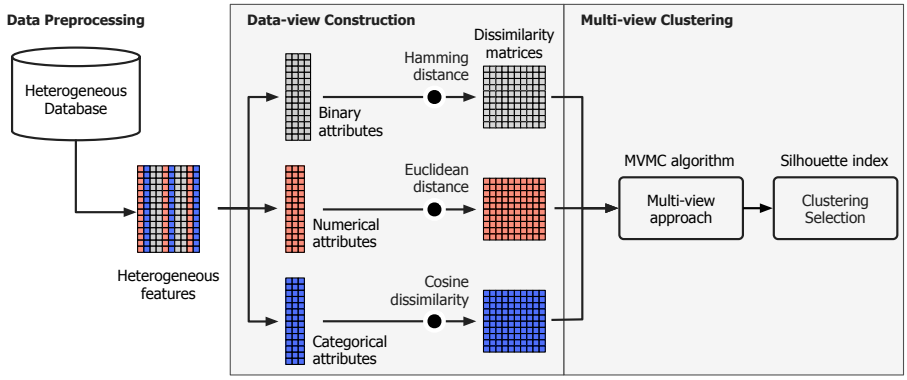

Fig. 1: Main stages and components of the proposed multi-view clustering methodology for a heterogeneous dataset.

unsupervised multi-view approach for clustering a heterogeneous database has been proposed and evaluated. This is because such approaches usually work on homogeneous data spliced across several datasets.

3.2 Multi-view Clustering Algorithm: `MVMC`

The `MVMC` algorithm is a multi-objective evolutionary approach to multi-view clustering that was developed to identify all optimal trade-offs between available data views [27]. It allows scalability to a significant number of views through the use of a many-objective optimizer. Specifically, `MVMC` uses a decomposition-based optimizer, MOEA/D [34], as the underlying search engine for its clustering approach. Furthermore, it employs a medoid-based representation, a representation that is more general than centroids, as it can be used both for problems defined in terms of feature spaces or dissimilarity matrices. In its current implementation, `MVMC` uses a fixed number of medoids, so requires the desired number of clusters as input.

`MVMC` focuses on a single cluster-quality criterion, but aims to optimize it concerning each view, resulting in a multi-objective optimization problem with as many clustering criteria as data views. Let $\mathbf{C}^r$ and $\mathbf{w}^r$ be the partition and weight vector, respectively, corresponding to the r-th subproblem. Also, let $\{D_1, \ldots, D_M\}$ denote M dissimilarity matrices, which represent M different data views and are each considered by a separate objective. `MVMC` then uses the within-cluster scatter as the optimization criterion, which, for the m-th objective of the r-th subproblem, is computed as:

$$f_m(\mathbf{C}^r) = \sum_{\mathbf{c}_k \in \mathbf{C}^r} \sum_{i,j \in \mathbf{c}_k} d_m(i,j) \ , \tag{1}$$

where $d_m(i,j)$ is the dissimilarity between the points i and j as defined in D_m.

`MVMC` overcomes one major dilemma of previous attempts at designing representations for multi-view clustering: how to ensure that these are scalable

without biasing the representation or decoding step toward one particular dissimilarity space. Specifically, the limitations of other representations are:

- For representations that are dissimilarity space agnostic, with each gene directly encoding cluster membership for each data point, the search space increases exponentially with the dataset size, affecting their scalability to large data.
- Representations that employ cluster prototypes in the form of centroids require the centroid to be represented in one or a concatenation of the feature spaces, which implies a single fixed weighting between views.
- Representations employing cluster prototypes (whether centroids or medoids) require a decoding step involving the assignment of data points to clusters. This step relies on using one or a sum of several dissimilarity functions, implying a single fixed weighting between views.

MVMC overcomes this issue by exploiting the availability of an explicit weight vector for each sub-problem in decomposition-based optimizers. Furthermore, employing a medoid-based encoding and accessing subproblem-specific weights in the decoding step avoids any prior bias towards one particular dissimilarity space whilst benefiting from a compact representation.

3.3 Selection of Clustering Solutions

The Silhouette index is often considered to be a more effective measure of cluster validity, as it combines both within and between-cluster variation of a partition. Unlike within-cluster scatter, maximizing the Silhouette index is potentially suitable for solution selection across a range of different numbers of clusters. For a given clustering solution $\mathbf{C}$ with N data points, the Silhouette index $\mathrm{Sil}(\mathbf{C})$ can be defined as the sum of individual Silhouette indexes $\{SW(i) \mid i = 1, ..., N\}$ [35]:

$$\mathrm{Sil}(\mathbf{C}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{SW}(i) = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{2}$$

where a_i represents the average distance from i to all other data points in its cluster. b_i represents the minimum distance of i to another cluster, where the distance between i and another cluster is calculated as the average distance from i to all data points in that cluster.

MVMC generates a set of non-dominated clustering solutions, but a single solution is usually required in practice. For this purpose, a model selection approach based on the Silhouette index is used [27]. This approach computes the index from a weighted dissimilarity matrix obtained from the weights assigned to the different data views during the clustering task.

4 Experimental Study

4.1 CHUL Database and Data-view Configurations

In this work, the different clustering methods were assessed and compared using the SSc patient database of the *Centre Hospitalier Universitaire de Lille*

(herein referred to as CHUL[7] database). The CHUL database was created in 2014 and held clinical information of 550 SSc patients with regular, detailed follow-up visits recorded on a standardized case-report form. Currently, the database contains more than 1500 patterns (patient visits) and nearly 400 attributes (e.g. demographic information, physical examination, laboratory exams, medical analyses). Two experienced clinicians (VS and DL authors) selected 39 relevant attributes, of which 22 are binary, 16 are numerical, and three are categorical (or nominal). In addition, data from the most recent visit of each patient were considered, limiting the analysis to 530 patterns. As a result, the clustering task was performed on 530 patterns described by 39 attributes with heterogeneous types. Three data views were generated from the CHUL database and used in the multi-view clustering algorithm:

- *Binary view*, {Bin}. This view is based on the binary dissimilarity data matrix computed with the Hamming distance on the 22 binary attributes.
- *Numerical view*, {Num}. This view is based on the numeric dissimilarity data matrix computed with the Euclidean distance on the 16 numerical attributes (integer and double data types) of the CHUL database.
- *Categorical view*, {Str}. This view is based on the categorical dissimilarity data matrix computed with the Cosine similarity measure on the 3 categorical attributes of the CHUL database.

For the MVMC algorithm, different view combinations of those data views were considered: {Bin,Num}, {Bin,Str}, {Num,Str}, and {Bin,Num,Str}. In addition, the {Num,Gower} configuration was considered, where the {Gower} view is a dissimilarity matrix created using the Gower distance from the union of the binary and categorical attributes in the CHUL dataset.

4.2 Reference Methods

To indicate baseline performance for the studied SSc data problem, we compare MVMC against two well-known and conceptually different clustering algorithms: K-medoids [33] and WARD hierarchical clustering method [38]. Our experiments apply K-medoids and WARD methods on four dissimilarity matrices, {HAM}, {EUC}, {COS}, and {GOWER}, using Hamming, Euclidean, Cosine, and Gower distances, respectively. These matrices were obtained from the entire CHUL dataset by transforming all attributes into numerical values.

The Silhouette scores obtained by WARD and K-medoids methods on each dissimilarity matrix were also computed, giving rise to possible comparisons between single-view and multi-view algorithms[8].

[7] SSc patients in the Internal Medicine Department of University Hospital of Lille, France, between October 2014 and December 2021 as part of the FHU PRECISE project (PREcision health in Complex Immune-mediated inflammatory diseaSEs); sample collection and usage authorization, CPP 2019-A01083-54.

[8] Note that the Silhouette score is intended to compare different partitions produced by a single method. Usually, the Rand index is preferred to the Silhouette score to compare two solutions when a ground-truth partition is available [35].

4.3 Parameter Settings

The settings for MVMC adopted in our experiments are as follows [27]. The population size is $NP = 100$, the number of generations is $G_{max} = 100$, the recombination probability is $Pr = 0.5$, the mutation probability is $Pm = 0.03$, and the neighborhood size is $T = 10$.

For the stochastic clustering methods analyzed and compared in this study, MVMC and K-medoids, a total of 31 independent executions were performed. In all cases, statistical significance is evaluated using the Kruskal–Wallis test, considering a significance level of $\alpha = 0.05$ and Bonferroni correction.

5 Results and Discussions

This section presents a series of experiments conducted on the CHUL dataset (530 patterns, 39 attributes) where different views and corresponding dissimilarity measures are considered according to attribute types. As described in Section 4.1, four dissimilarity matrices and five data-view configurations are used by two single-view, WARD and K-medoids algorithms, and the multi-view approach MVMC.

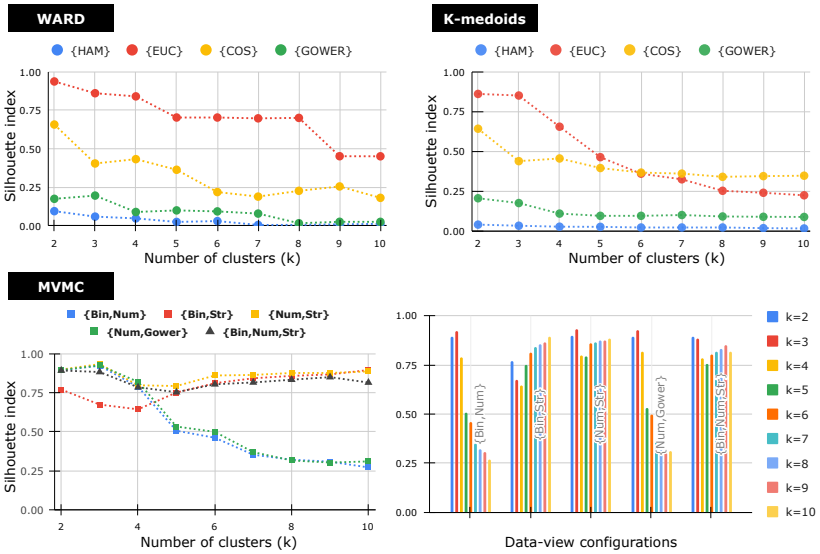

Fig. 2: Illustration of the clustering performance obtained by the different algorithm configurations when varying the number of clusters, $K = \{k | 2 \leqslant k \leqslant 10\}$.

5.1 Clustering Performance

This first experiment aims to analyze the clustering performance of the clustering algorithms with the number of clusters. Thus, the results obtained by WARD and K-medoids will serve as a reference (baseline) when compared with those obtained by the multi-view approach, MVMC. The WARD and K-medoids algorithms

Table 1: Clustering performance in terms of the Silhouette index obtained by the different algorithm configurations when varying k, $K = \{k | 2 \leqslant k \leqslant 10\}$. The best Silhouette value scored for each algorithm configuration has been shaded and highlighted in bold and, additionally, the statistically best ($\alpha = 0.05$) results are highlighted in boldface.

Alg.	Data views	Number of clusters (k)								
		$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
WARD	{HAM}	**0.095**	0.060	0.048	0.025	0.031	0.006	0.006	0.010	0.013
	{EUC}	**0.937**	**0.861**	0.840	0.703	0.703	0.698	0.701	0.452	0.451
	{COS}	**0.657**	0.405	0.433	0.364	0.219	0.190	0.227	0.256	0.182
	{GOWER}	**0.175**	**0.196**	0.090	0.100	0.094	0.079	0.017	0.026	0.026
K-medoids	{HAM}	**0.043**	0.036	0.030	0.029	0.025	0.025	0.025	0.021	0.019
	{EUC}	**0.861**	**0.851**	0.656	0.465	0.362	0.327	0.255	0.242	0.227
	{COS}	**0.644**	0.440	0.457	0.397	0.369	0.362	0.342	0.347	0.349
	{GOWER}	**0.208**	**0.178**	0.112	0.098	0.098	0.103	0.094	0.092	0.091
MVMC	{Bin,Num}	**0.894**	**0.922**	0.787	0.506	0.461	0.352	0.321	0.308	0.271
	{Bin,Str}	0.770	0.674	0.643	0.751	**0.811**	**0.842**	**0.857**	**0.867**	**0.895**
	{Num,Str}	**0.898**	**0.933**	0.797	0.793	**0.860**	**0.863**	**0.876**	**0.876**	**0.886**
	{Num,Gower}	**0.895**	**0.925**	0.819	0.533	0.499	0.369	0.316	0.302	0.310
	{Bin,Num,Str}	**0.892**	**0.891**	0.785	0.753	**0.803**	**0.815**	**0.833**	**0.849**	**0.815**

were used to separately cluster the four dissimilarity matrices {HAM}, {EUC}, {COS}, and {GOWER}, whereas MVMC used five different data-views combinations.

The experiment was conducted as follows. First, for each clustering algorithm and each data view, a collection of $\mathbb{C}$ partitions were generated by varying the number of clusters k in the range $K = \{k | 2 \leqslant k \leqslant 10\}$. Then, in a second step, each clustering solution in collection $\mathbb{C}$ was evaluated using the Silhouette index. Usually, the partition(s) with the best index values are considered the final solutions that best fit the data problem. This procedure is commonly used when the number of clusters is unknown and needs to be determined using a cluster validity index. For this purpose, the Silhouette index is well known and has performed satisfactorily in practice [24]. The results of this analysis are summarized in Figure 2, with more detailed results, and their statistical significance, presented in Table 1.

The average Silhouette index values tend to decrease as the number of clusters increases from two to ten for traditional single-view algorithms, i.e., the Silhouette index suggests that the most appropriate number of clusters is at the beginning of the range of explored clusters. Then, it is observed that the index quickly loses its discriminative ability to find other suitable underlying structures in this highly heterogeneous dataset. Moreover, this monotonous decreasing convergence behavior is observed in both single-view algorithms and is independent of the type of proximity measure used in the experiments.

On the other hand, regarding the clustering performance obtained by the multi-view clustering algorithm MVMC using the five data-view configurations, it is observed that in general, (i) the algorithm obtained higher average Silhouette values than traditional clustering approaches and (ii) that the Silhouette values are changing as the number of clusters increases (i.e. the values increase and decrease). In addition, two types of Silhouette convergences are observed concerning the performance of the different data configurations. First, configurations {Bin,Num} and {Num,Gower} obtained very similar convergence

results: they start by slightly increasing, up to a certain k, and then start to decrease as the number of clusters increases further. Second, for data configurations {Bin,Str}, {Num,Str}, and {Bin,Num,Str} the Silhouette values increase, decrease, increase again to a certain threshold, and then remain constant. These Silhouette index fluctuations indicate that multiple suitable cluster structures are encountered across the range of explored clusters. Thus, in the following subsection, we investigate the selection of the most appropriate clustering solutions.

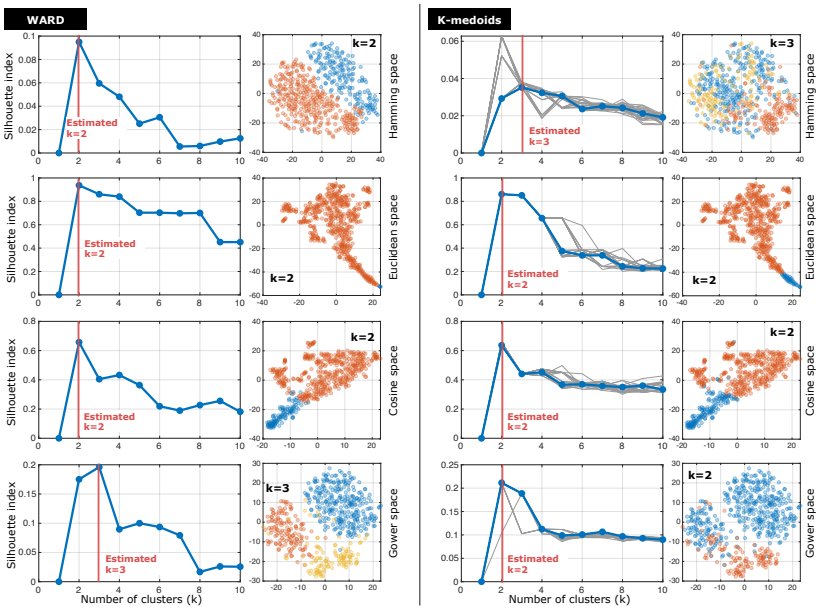

Fig. 3: Best clustering solutions obtained by WARD (left) and K-medoids (right) algorithms using the Silhouette index. For each subfigure, the median convergence plot is shown in blue. The best solution is marked in red. The corresponding clustering solution is visualized in the embedding space associated with a proximity measure.

5.2 Selection of Clustering Solutions

An important problem in cluster analysis is to determine the number of clusters from the inherent information in a clustering structure [23]. Thus, the following experiment aims to find both the most appropriate number of clusters and its corresponding clustering solution from a collection of solutions using the Silhouette index. This experiment was conducted as follows. First, the solutions(s) with the highest Silhouette value(s) are selected among the collection of solutions generated by a clustering algorithm. Subsequently, the chosen solution(s) is visualized in an embedded two-dimensional feature space, obtained from a dissimilarity matrix using the t-SNE [32] projection technique (parameters: *n_components*=2, *n_iter*=100, *perplexity*=30). The resulting clustering solutions of this analysis are presented in Figures 3-4.

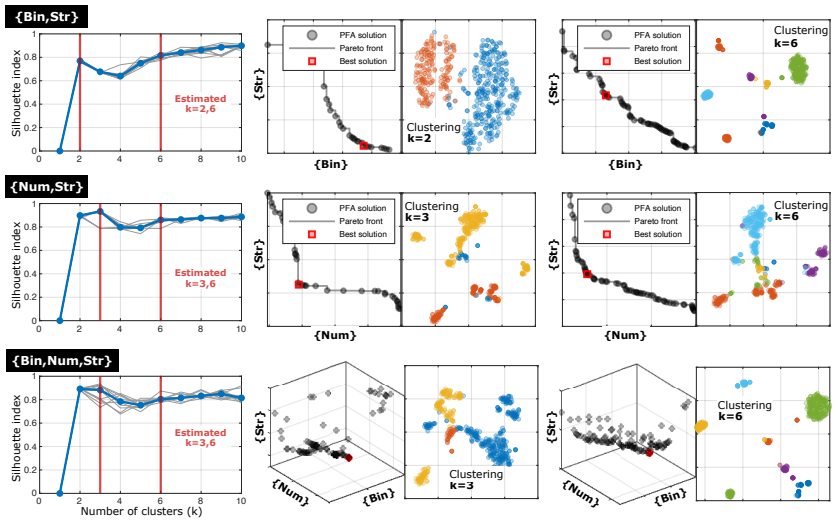

Fig. 4: `MVMC` clustering solutions for the configurations, `{Bin,Str}`, `{Num,Str}`, and `{Bin,Num,Str}`. Each configuration includes the convergence plots shown in blue and gray, with the two best solutions marked red. Then, for each selected solution, (i) the Pareto front approximation (PFAs) and (ii) the clustering solution, which is visualized in a weighted embedding space associated with the data views in the configuration.

Figure 3 presents the selected solutions for the two single-view algorithms. In general, we can observe that the choice of the distance function over the original heterogeneous dataset considerably influences the two-dimensional distribution of t-SNE projections. Furthermore, there is a clear tendency for the Silhouette index to discover two clusters in most scenarios, except for configurations $\texttt{WARD}_{\texttt{\{GOWER\}}}$ and $\texttt{K-medoids}_{\texttt{\{EUC\}}}$, where the number of groups is three.

Regarding the clustering solutions generated by the multi-view approach `MVMC`, from Figure 5 (Appendix), it is clear that the determined number of clusters is three as the Silhouette index obtained its highest point value at this point, $k = 3$. Figure 4 illustrates the generated clustering solutions for the data-view configurations, `{Bin,Str}`, `{Num,Str}`, and `{Bin,Num,Str}`. Two solutions with the best Silhouette values were chosen for each configuration in this scenario. Firstly, we observe that the best clustering solutions tend to be found at the knee of the Pareto front approximations (PFAs), red box in the PFA, representing trade-offs between the views involved. These compromise points suggest that the consensus clustering solution exploits pieces of information from all the multiple data views in a complementary manner. As a result, the multi-view clustering setting reveals three and six clusters (inflection points in convergence plots). Interestingly, the combination of the (mixed) data-view contributions produces embedded feature spaces with observable groups, particularly for the six-cluster solutions, as illustrated in Figure 4.

Finally, Table 2 presents two clustering solutions (**P** and **G**) obtained by the `MVMC` algorithm with the data-view configuration `{Bin,Num,Str}`. The first clustering solution contains two clusters and is shown in the first two columns

Table 2: Two final clustering solutions obtained by MVMC with {Bin,Num,Str}.

Descriptive Atts.[1]	P(k = 2)		G(k = 6)					
	P1	P2	G1	G2	G3	G4	G5	G6
Cluster Size	177	353	255	70	68	50	50	37
Sex (m,f)	(25,75)	(12,88)	(10,90)	(29,71)	(13,87)	(38,62)	(18,82)	(14,86)
SSc Type (dc,lc,sc)	(40,59,1)	(10,72,18)	(0,82,18)	(29,69,3)	(13,87,0)	(92,8,0)	(0,66,34)	(81,19,0)
Active DU (y,n)	(60,40)	(42,58)	(41,59)	(54,46)	(53,47)	(68,32)	(36,64)	(65,35)
Active SRC (y,n)	(3,97)	(0,100)	(0,100)	(4,96)	(0,100)	(4,96)	(0,100)	(0,100)
ILD (y,n)	(98,2)	(6,94)	(1,99)	(100,0)	(85,15)	(100,0)	(24,76)	(3,97)
PH (y,n)	(12,88)	(8,92)	(7,93)	(11,89)	(15,85)	(16,84)	(6,94)	(11,89)
Calcinosis (y,n)	(10,90)	(13,87)	(14,86)	(6,94)	(18,82)	(6,94)	(4,96)	(19,81)
Joint Sx (y,n)	(34,66)	(41,59)	(40,60)	(31,69)	(37,63)	(34,66)	(42,58)	(43,57)
Intestinal Sx (y,n)	(27,73)	(30,70)	(31,69)	(23,77)	(32,68)	(28,72)	(16,84)	(43,57)
mRSS	8.78±7.6	5.73±4.8	4.30±3.3	8.58±8.2	7.05±6.0	11.28±7.4	5.63±5.7	10.47±6.2
LVEF	63.44±28.6	64.74±23.3	63.91±5.5	60.85±4.4	65.05±4.5	62.84±6.9	61.20±6.5	65.06±5.4
FVC	87.41±27.1	102.13±29.4	107.83±19.1	83.49±23.6	101.95±16.3	85.54±24.1	106.07±20.4	103.57±21.9
DLCO	55.54±16.5	69.38±21.9	74.21±22.0	54.78±18.8	68.08±18.0	56.07±19.4	73.84±19.1	70.92±17.1
Score EUSTAR	1.70±1.5	1.55±1.3	1.42±1.1	1.59±1.3	1.77±1.5	2.38±1.8	1.61±1.3	2.32±1.6
Score Medsger	1.41±0.8	1.25±0.7	1.46±0.8	1.67±0.9	1.77±0.9	1.67±0.8	1.71±1.0	2.17±1.2

[1] Sex: m (male), f (female); SSC Type: dc / lc (diffuse / limited cutaneous), sc (sine scleroderma); DU: digital ulceration; SRC: scleroderma renal crisis; ILD: interstitial lung disease; PH: pulmonary hypertension; Sx: symptoms; mRSS: mean Rodnan skin score; LVEF: left ventricular injection fraction; FVC: forced vital capacity; DLCO: diffusion lung capacity for carbon monoxide; EUSTAR: european scleroderma trials and research.

in gray. In contrast, the second solution involves six groups and is described in the last six columns in light blue. Regarding clinical relevance, solution **P** exhibits two groups of patients separated on the basis of the presence of ILD, and interestingly not regarding the cutaneous involvement (historical subclassification [37]). The six-cluster solution provided a better delineation of six homogeneous groups, which best captured the patients' variability in terms of the disease severity as expressed by the EUSTAR and Medsger scores. G1 included the majority of patients with mild disease. G4 and G6 were mostly patients with diffuse cutaneous involvement. G2, G3, and G4 were patients with ILD and different degrees of severity as shown by the FVC and DLCO values. PH was found with a high prevalence in G2, G3, G4 and G6, but DLCO values unveiled that G2 and G4 were the most severe regarding gas exchange capacity.

6 Conclusion

This work explores the benefits of multi-view clustering to identify groups of systemic sclerosis (SSc) patients, a highly heterogeneous auto-immune disease, within electronic health records (EHRs) capturing several types of attributes. Our approach avoids the premature integration of attribute types before cluster analysis through a multi-objective evolutionary algorithm called MVMC. MVMC integrates multiple data types into the clustering process in the form of data views, explores trade-offs between them, and determines consensus clusters supported across these views. This comprehensive classification integration of multiple and various data sources helped to discover meaningful clustering solutions ($\mathbf{P}_{k=2}$ and $\mathbf{G}_{k=6}$) that will help to better understand disease complications and treatment goals.

Acknowledgments The authors are grateful to the University of Lille, CHU Lille, and INSERM, founded by the MEL through the I-Site cluster humAIn@Lille.

Appendix

This Appendix includes figures complementing the results of the experiments presented in Section 5. From Figure 5 (Appendix), it is clear that the determined number of clusters is three as the Silhouette index obtained its highest point value at this point, $k = 3$. Also, from the Pareto front approximations obtained by these configurations, a substantial inference of the {Num} view is observed over the {Bin} and {Gower} views, respectively. Accordingly, the clustering solutions and the weighted embedding space are remarkably similar between these two data-view configurations.

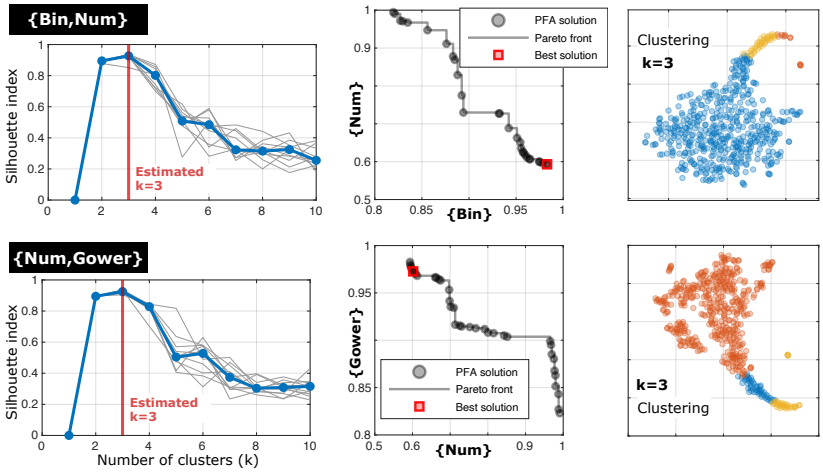

Fig. 5: MVMC clustering solutions for two data-view configurations, {Bin,Num} and {Num,Gower}. Each configuration includes (i) the convergence plots shown in blue and gray, with the best solution marked red; (ii) the Pareto front approximation corresponding to the estimated k value; (iii) the clustering solution, which is visualized in a weighted embedding space associated with the data views in the configuration.

References

1. Abdullin, A., Nasraoui, O.: Clustering heterogeneous data sets. In: American Web Congress. pp. 1–8. IEEE (2012)
2. Ahmad, A., Dey, L.: A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering **63**(2), 503–527 (2007)
3. Ahmad, A., Khan, S.S.: Survey of state-of-the-art mixed data clustering algorithms. IEEE Access **7**, 31883–31902 (2019)
4. Ahmad, A., Khan, S.S.: initkmix-a novel initial partition generation algorithm for clustering mixed data using k-means-based clustering. Expert Systems with Applications **167**, 114149 (2021)
5. Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., Cremers, D.: Clustering with deep learning: Taxonomy and new methods. arXiv:1801.07648 (2018)
6. Banfield, J.D., Raftery, A.E.: Model-based gaussian and non-gaussian clustering. Biometrics **49**(3), 803–821 (1993)

7. Basel, A.J., Rui, F., Nandi, K.A.: Integrative cluster analysis in bioinformatics. John Wiley & Sons (2015)
8. Bécue-Bertaut, M., Pagés, J.: Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computational Statistics & Data Analysis **52**(6), 3255–3268 (2008)
9. Ben Ali, B., Massmoudi, Y.: K-means clustering based on gower similarity coefficient: A comparative study. In: International Conference on Modeling, Simulation and Applied Optimization (ICMSAO). pp. 1–5. IEEE (2013)
10. Budiaji, W., Leisch, F.: Simple k-medoids partitioning algorithm for mixed variable data. Algorithms **12**(9) (2019)
11. de Carvalho, F., Lechevallier, Y., de Melo, F.M.: Partitioning hard clustering algorithms based on multiple dissimilarity matrices. Pattern Recognition **45**(1), 447–464 (2012)
12. de Carvalho, F.d.A., Lechevallier, Y., de Melo, F.M.: Partitioning Hard Clustering Algorithms based on Multiple Dissimilarity Matrices. Pattern Recognition **45**(1), 447–464 (2012)
13. Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C.: A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 263–268. KDD '01, Association for Computing Machinery, New York, NY, USA (2001)
14. de Carvalho, F., Lechevallier, Y., Despeyroux, T., de Melo, F.M.: Multi-view clustering on relational data. In: Zighed, F., Abdelkader, G., Gilles, P., Venturini, B.D. (eds.) Advances in Knowledge Discovery and Management, pp. 37–51. Springer (2014)
15. Foss, A.H., Markatou, M., Ray, B.: Distance metrics and clustering methods for mixed-type data. International Statistical Review **87**(1), 80–109 (2019)
16. Fraley, C., Raftery, A.E.: How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal **41**(8), 578–588 (01 1998)
17. Green, P.E., Rao, V.R.: A note on proximity measures and cluster analysis. Journal of Marketing Research **3**(6), 359–364 (1969)
18. Harikumar, S., PV, S.: K-medoid clustering for heterogeneous datasets. Procedia Computer Science **70**, 226–237 (2015)
19. Hsu, C.C., Chen, C.L., Su, Y.W.: Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences **177**(20), 4474–4492 (2007)
20. Huang, J., Ng, M., Hongqiang Rong, Zichen Li: Automated variable weighting in k-means type clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(5), 657–668 (2005)
21. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: The Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 21–34 (1997)
22. Hunt, L., Jorgensen, M.: Clustering mixed data. WIREs Data Mining and Knowledge Discovery **1**(4), 352–361 (2011)
23. José-García, A., Gómez-Flores, W.: Automatic clustering using nature-inspired metaheuristics: A survey. Applied Soft Computing **41**, 192–213 (2016)
24. José-García, A., Gómez-Flores, W.: A survey of cluster validity indices for automatic data clustering using differential evolution. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 314–322. ACM Press (2021)
25. José-García, A., Handl, J.: On the interaction between distance functions and clustering criteria in multi-objective clustering. In: International Conference on Evolutionary Multi-Criterion Optimization. pp. 504–515. Springer (2021)

26. José-García, A., Handl, J., Gómez-Flores, W., Garza-Fabre, M.: Many-view clustering: An illustration using multiple dissimilarity measures. In: Genetic and Evolutionary Computation Conference - GECCO '19. pp. 213–214. ACM Press, Prague, Czech Republic (2019)
27. José-García, A., Handl, J., Gómez-Flores, W., Garza-Fabre, M.: An evolutionary many-objective approach to multiview clustering using feature and relational data. Applied Soft Computing **108** (2021)
28. Landi, I., Glicksberg, B.S., Lee, H.C., Cherng, S., Landi, G., Danieletto, M., Dudley, J.T., Furlanello, C., Miotto, R.: Deep representation learning of electronic health records to unlock patient stratification at scale. npj Digital Medicine **3**(1), 96 (2020)
29. Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering **14**(4), 673–690 (2002)
30. Liu, C., Chen, Q., Chen, Y., Liu, J.: A fast multiobjective fuzzy clustering with multimeasures combination. Mathematical Problems in Engineering **2019**, 1–21 (2019)
31. Liu, C., Liu, J., Peng, D., Wu, C.: A general multiobjective clustering approach based on multiple distance measures. IEEE Access **6**, 41706–41719 (2018)
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(11) (2008)
33. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297. University of California Press (1967)
34. Qingfu Zhang, Hui Li: MOEA/D: A multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on Evolutionary Computation **11**(6), 712–731 (2007)
35. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
36. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y.: A comparison study on similarity and dissimilarity measures in clustering continuous data. PLOS ONE **10**(12) (2015)
37. Sobanski, V., Giovannelli, J., Allanore, Y., et al.: Phenotypes determined by cluster analysis and their survival in the prospective european scleroderma trials and research cohort of patients with systemic sclerosis. Arthritis & Rheumatology **71**(9), 1553–1570 (2019)
38. Theodoridis, S., Koutrumbas, K.: Pattern Recognition. Elsevier Inc., fourth edn. (2009)
39. Vandromme, M., Jacques, J., Taillard, J., Jourdan, L., Dhaenens, C.: A biclustering method for heterogeneous and temporal medical data. IEEE Transactions on Knowledge and Data Engineering **34**(2), 506–518 (2022)
40. van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. WIREs Computational Statistics **11**(3), e1456
41. Wei, M., Chow, T., Chan, R.: Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. Entropy **17**(3), 1535–1548 (2015)