The Clustering Analysis on Public Health Data with Missing Values Based on Dimension Reduction Methods

Honghao Zhao(✉), Weiyi Ding, Fang Ye, Weimeng Yuan, and Hangyu Chen

Department of Decision Sciences, School of Business, Macau University of Science and Technology, Taipa, Macao, China
hhzhao@must.edu.mo

Abstract. With the development of medical information digitization, machine learning techniques have become a popular method of mining medical health data for hidden information and knowledge. Health data from normal medical checking is usually limited. However, public health data from magnetic resonance (MR) are usually high-dimensional data with missing values. This paper presents a clustering analysis of such health data with a series of steps, including filling missing values, dimension reduction, and clustering to provide a framework with potential solutions to the problems of missing value and high data dimension. Our results show that the UMAP method is the most effective one for dimension reduction, and the K-means clustering method works well in most cases.

Keywords: Public Health Data · Clustering Analysis · Matrix Completion · Dimension Reduction

1 Introduction

People became more concerned about their quality of life and healthcare with the development of society. The usage of electronic medical records has increased the frequency of physical examinations. Health testing data is commonly referred to as medical data, and due to the numerous test subjects and consistent testing frequency, it has become an essential component of big data in medicine. The efficiency use of big data and machine learning technologies that can uncover hidden information in explosive data, which support the design of health insurance plans by government organizations, and medical research at institutes.

Due to the challenges of gathering consistently labeled data and extracting knowledge from complex supervised learning, unsupervised learning techniques are gaining popularity. Among them, clustering methods can greatly simplify data models and retain insightful knowledge from massive data. For instance, the knowledge extracted by clustering in disease research can be applied to disease prevention, treatment methods, and novel drug discovery [1]. In the earlier 21st century, cluster analysis as a research technique regained attention in the fields of bioinformatics and health [2, 3].

To achieve clustering, the method itself relies on the similarity among data samples and necessitates relatively poor data quality. However, this may become challenge, when

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 45–51, 2024.
https://doi.org/10.2991/978-94-6463-242-2_7

the data is high-dimensional with missing values, which is commonly in public health records. Data incompleteness can be ascribed by several factors. When a disease is diagnosed, various patients have varying degrees of severity and thus obtain varying diagnosis results, which may give rise to a lack of certain characteristics in data sample. The absence of medical health data is further influenced by the consistency of testing procedures and privacy concerns regarding data disclosure.

For data with high dimension, unsupervised dimensionality reduction algorithms, which apply clustering algorithms to the deducted dimensional space, are a typical solution for the clustering for high-dimensional data. Unsupervised dimensionality reduction can be broken down into two parts: feature extraction using low-dimensional embedding search techniques, and feature selection [4]. The data can also be visualized if the dimensionality is only 2 or 3 dimensions.

In this paper, we intend to provide a framework for clustering analysis of healthcare data with potential solutions to the problems of missing value and high data dimension. In particular, the filling of missing value is achieved by matrix completion method, and various dimension reduction methods are compared. Finally, the common clustering methods are investigated based on the optimal dimension reduction results.

2 Description of Methods

In this section we describe the major methods used in this research. These methods are implemented to handle different problems when analysis public health data, including fixing missing values, dimension reduction and clustering methods.

2.1 Matrix Completion Algorithm

The idea behind matrix completion, also known as low-rank matrix recovery. An early idea was to fill in the missing part of X by using a low-rank matrix Z with no missing values through solving the following problem,

$$min\, rank(Z), s.t. \sum\nolimits_{(i,j)\in\Omega} \left(Z_{ij} - X_{ij}\right) \leq \eta \tag{1}$$

Due to the NP-hard problem complexity, the SVT algorithm was proposed to use the kernel parametrization of ||Z|| as an approximation optimization goal [5].

2.2 Dimension Reduction

Dimensionality reduction methods can retain the most important features and denoise the data. We discuss 4 dimensionality reduction methods, which are Principal component analysis (PCA), Multidimensional Scaling (MDS), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP).

PCA is to retain as much variation in the original dataset as possible. PCA can compress and denoise data in addition to preventing curse of dimensionality and displaying high-dimensional data in an understandable manner [6]. The MDS is to maintain as much distance in low-dimensional space between samples of original data as possible. It

can clearly identify the aggregation of high-dimensional samples [7]. T-SNE technique gives more priority on conserving the probability distribution between data samples, resolving the SNE crowding issue [8]. The Uniform Manifold Approximation and Projection (UMAP), performs well in retaining the original probability distribution. This technique reduces dimensions by searching for low-dimensional projections of the data with the closest equivalent fuzzy topology, which is based on the framework of Riemannian geometry and algebraic topology and presupposes that the data are equally distributed across Riemannian manifolds [9].

2.3 Clustering Methods

In this study, the reduced-dimensional data are clustered and analyzed through three common clustering algorithms: K-Means clustering, spectral clustering, and hierarchical clustering, which are introduced below.

K-Means clustering, one of the prototype-based clustering methods which identifies the representative prototypes of each cluster by initializing and solving the prototypes iteratively. The goal is to minimize the intra-class differences between clusters. E is the optimization objective of K-Means clustering in Eq. (2).

$$E = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2 \tag{2}$$

Given the high complexity of brute force, it can be equivalently transformed to the optimization goal E using the prototype vector. A straightforward algorithm can be employed to find a locally optimal solution, which is given by,

$$E = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{ij} - \mu_i \right)^2 \mu_i = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij} \tag{3}$$

The algorithm's pseudocodes are listed below.

Step 1: Choose k data points at random to serve as the prototype vectors.

Step 2: Determine the separation between each sample and the prototype vector, then place each sample in the group with the closest prototype vector.

Step 3: Update the vectors by recalculating them in accordance with Eq. (3).

Step 4 Repetition of steps 2 through 3 will stop updating the prototype vectors.

Spectral clustering is a graph theory-based approach [10] addressing the issue of cut graphs in undirected graphs. The similarity between samples is presented by the weight of the connection between points on the undirected graph. More weights are assigned to points with higher similarity, and vice versa.

Hierarchical clustering provides a succession of clustering for merging or splitting a dataset by identifying the merging method that optimizes the target distance in each iteration and defining a target distance to achieve the dataset clustering. Each sample in the dataset is considered as a separate cluster initially, and AGNES (Agglomerative Nesting) selects the two clusters that are the closest to one another to merge. The distance is usually given by,

$$dist_{min}\left(C_i, C_j\right) = min_{a \in C_i, b \in C_j}, dist(a, b) \tag{4}$$

Table 1. Hopkins statistics for different dimension reduction methods

Downscaling methods	Hopkins statistics
UMAP	0.9273
MDS	0.9149
t-SNE	0.8846
PCA	0.9121
Raw data	0.8699

3 The Experiment and Results

3.1 The Data Set and Filling Missing Values

The data of this study is collected from a health screening technique called micro-magnetic resonance (MR), which uses tiny wearable sensors to send signals into the body and then receive signals back from the body. The dataset contains 2,170 records and 279 dimensions with numbers of missing values. In this study, the missing value dimensions in the data were filled using the SVT method in the matrix complementation.

3.2 Results of Dimension Reduction

The Hopkins statistics is most frequently utilized to determine if the data are randomly and evenly distributed in clustering trend analysis, which is given by

$$\mathrm{H} = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} v_i} \tag{5}$$

where u_i and v_i present the distance between the closest sample in B and the sample drawn at random, in different sets.

As shown in Table 1, the UMAP method has the largest Hopkins statistics, which indicates a better dimension reduction result. Similar results can also be observed in Fig. 1, which indicates that the UMAP method is more suitable in this case for the purpose of dimension reduction.

3.3 Clustering Results

In this study, we set the number of clusters to be between 2 and 10, use various clustering algorithms to group the data into smaller groups after UMAP is reduced to 2 dimensions, and use unsupervised clustering evaluation metrics to compare the models for the various clustering algorithms' outputs in order to choose the most appropriate clustering model for the dataset. In particular, three common clustering algorithms—K-Means clustering, spectral clustering, and hierarchical clustering—are used to cluster the data. Furthermore, the silhouette coefficient (SC), Davies Bouldin index (DB), Calinski harabasz index (CH) and Model Order Estimator (MO) are employed to assess the effectiveness of clustering algorithm outcomes.

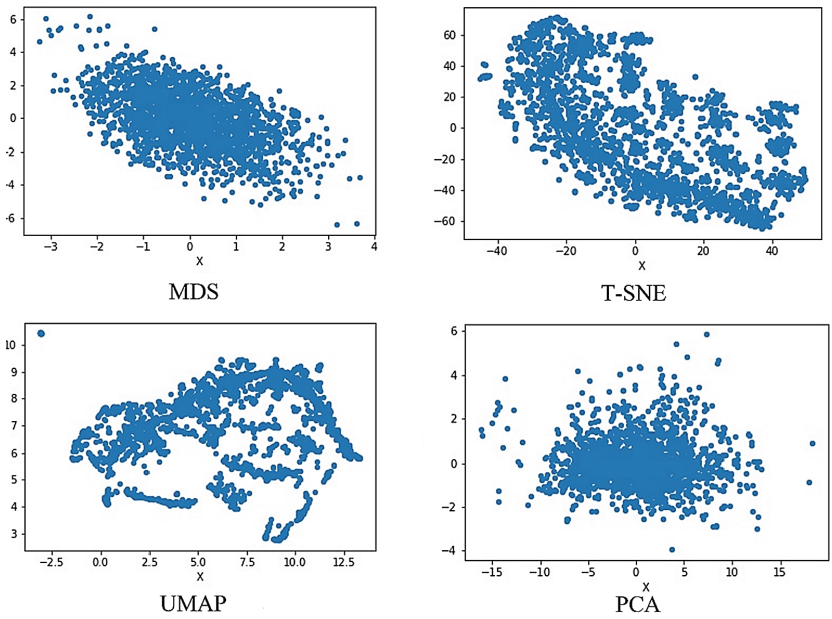

Fig. 1. Graphical representation of downscaling results for various methods

The four subplots of Fig. 2 display the clustering results associated with the various assessment indicators. In particular, the subfigures on the upper-left corner, upper-right corner, lower-left corner, and lower-right corner represent the results for CH, SC, DB and MO, respectively. Various clustering methods employ various legend labels: Circles are used in K-Means clustering, triangles are used in spectral clustering, and squares are used in hierarchical clustering. The quality of the clusters increases with increasing SC and CH indicators, and decreases with decreasing DB and Model order Estimator.

Table 2 presents the optimal number of clusters and their corresponding indicator values. The evaluation indicators' upper and lower arrows show the positive and negative correlations between the indicators and the clusters' quality. Table 2 demonstrates that K-Means clustering outperforms other clustering methods overall in most evaluation metrics, expect that Spectral clustering performs the best in terms of DB index.

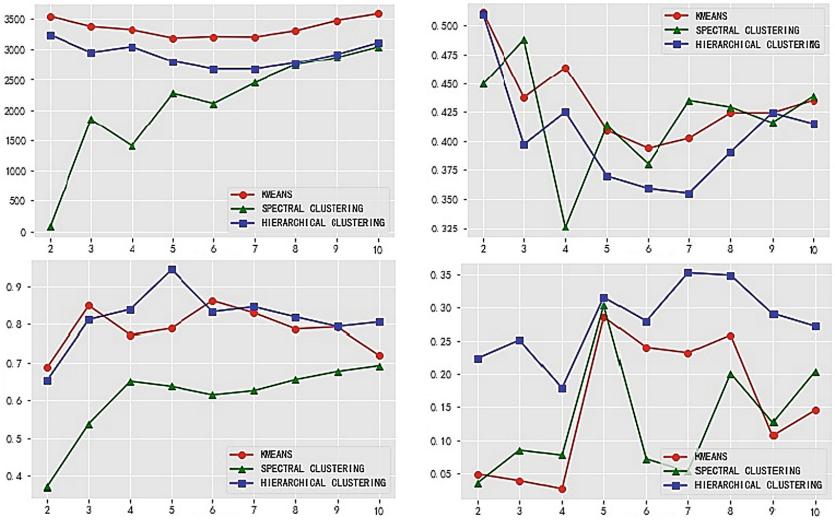

Fig. 2. UMAP dimensionality reduction-clustering evaluation results

Table 2. Optimal number of clusters (ONC) and Indicator values (IV)

	K-Means		Spectral clustering		Hierarchical clustering	
Indicator	ONC	IV	ONC	IV	ONC	IV
SC ↑	**2**	**0.511**	3	0.488	2	0.510
CH ↑	**10**	**3581.798**	10	3036.681	2	3240.201
DB ↓	2	0.687	**2**	**0.371**	2	0.653
MO↓	**4**	**0.027**	2	0.036	4	0.179

4 Conclusions

Medical data has gradually become prevalent in data mining research recently due to the pandemic. In order to overcome the frequent issues of high dimensionality, the presence of a significant number of missing values, and the absence of labels, a novel type of health testing data is chosen in this article. A collection of clustering analysis procedures is presented in this study for the clustering analysis.

The results show that the UMAP method is most effective in terms of dimension reduction. K-means clustering outperforms in most evaluation criteria, except that the spectral clustering performs well in DB index. To simplify the clustering, we only consider setting the number of clusters for a clustering algorithm to 2–10, and the practical implication of the results are not discussed, which are worthy of further research.

References

1. Xu, R., & Wunsch, D. C. (2010). Clustering algorithms in biomedical research: a review. IEEE reviews in biomedical engineering, 3, 120-154.
2. Wosiak, A., & Zakrzewska, D. (2018). Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. Complexity, 2018.
3. Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. Neural computation, 16(6), 1299-1323.
4. Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, 54(5), 3473-3515.
5. Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on optimization, 20(4), 1956-1982.
6. Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
7. Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. Measurement, judgment and decision making, 179–250.
8. Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. Advances in neural information processing systems, 15.
9. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
10. Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4), 395-416.