Hierarchical Clustering for Mixed Feature-Type Complex Data

Huiwen Wang
School of Economics and Management, Beihang University, Beijing 100191, China - wanghw@vip.sina.com
Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beijing 100191, China

Cheng Wang*
School of Economics and Management, Beihang University, Beijing 100191, China - wangcheng@buaa.edu.cn
Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beijing 100191, China

Yuan Wei
School of Economics and Management, Beihang University, Beijing 100191, China - weiyuan1143@163.com
Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beijing 100191, China

Abstract
With the rapid development of cross-platform data collection technology and the coming of the big data era, there are always a mixture of single-valued data, symbolic data, composition data and functional data in one table, which can be called mixed feature-type complex data. So far no operating rules among these different types of data have been given, thus no multivariate models based on mixed data can be built, which has already become a major obstacle of data analysis in economics and management. In many real applications that involve mixed feature-type complex data, clustering is needed to classify the information. In this paper, we present dissimilarity matrix on the multidimensional mixed feature-type complex space to measure the distance between different observations, and propose agglomerative hierarchical clustering method for mixed feature-type complex data based on the dissimilarity matrix. Experiments with simulated data in a framework of a Monte Carlo schema are proposed to verify the performance of the hierarchical clustering method based on the dissimilarity matrix.

Keywords: Dissimilarity Matrix; Symbolic Data; Composition Data; Functional Data.

1. Introduction
Nowadays, lots of companies, banks, financial markets, government agencies and electronic commercial enterprises have developed massive and untapped cross-platform data with the long-term accumulation. These datasets often have a huge number of observations and variables of high dimension, which can be both quantitative and qualitative, and some of these data are even collected at extremely high frequency. How to efficiently explore the information in large-scale data and use it in diagnosing and forecasting the system state, has become a new opportunity and challenge in the field of economics and management.

In current research, there are always a mixture of single-valued data, symbolic data (Diday, 1987), composition data (Aitchison, 1982) and functional data (Ramsay, 1982) in one table, which can be called mixed feature-type complex data. So far no operating rules among different types of data have been given, thus no multivariate models based on mixed data can be built, which has already become a major obstacle of data analysis in economics and management. In many real applications that involve mixed feature-type complex data, clustering is needed to classify the information. Clustering has emerged as a popular technique for pattern recognition, artificial intelligence, image processing, data mining and so on. Clustering methods seek to organize a set of items into clusters such that items within a given cluster have a high degree of similarity, whereas those of different clusters have a high degree of dissimilarity (Jain et al., 1999). The most popular clustering techniques are hierarchical and partitioning methods (Gordon, 1999). Hierarchical methods yield complete hierarchy, while partitioning methods seek to obtain a single partition of the input data into a fixed number of clusters.

Hierarchical methods can be agglomerative or divisive (Gordon, 1999). Agglomerative methods yield a sequence of nested partitions starting with trivial clustering in which each item is in a unique cluster and ending with trivial clustering in which all items are in the same cluster. A divisive method starts with all items in a single cluster and performs a splitting procedure until a stopping criterion is met (usually upon obtaining a partition of singleton clusters).

In symbolic data analysis, clustering methods differ in the type of the considered symbolic data, in their cluster struc-
In compositional data analysis, Martín-Fernández et al. (1998) introduced different measures of difference for compositional data. Hierarchical clustering algorithm by introducing the concept of mutual dissimilarity value is proposed for clustering of each cluster in the hierarchy. Guru and Kiranagi (2005) proposed a novel dissimilarity measure and a modified data sets containing numeric and symbolic feature values. Chavent (2000) proposed a divisive clustering method for mixed feature variables and presents dendrograms obtained from the application of standard linkage methods for agglomeration based on minimum dissimilarity. Ichino and Yaguchi (1994) defined generalized Minkowski metrics for symbolic data that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterization of each cluster in the hierarchy. Guru and Kiranagi (2005) proposed a novel dissimilarity measure and a modified hierarchical clustering algorithm by introducing the concept of mutual dissimilarity value is proposed for clustering symbolic patterns. Ippino and Verde (2006) used Wasserstein metric for an agglomerative hierarchical clustering of histogram data based on the Ward criterion.

In compositional data analysis, Martín-Fernández et al. (1998) introduced different measures of difference for compositional data and hierarchical clustering methods. The distances verify the requirements of scale invariance, permutation invariance, perturbation invariance and sub-compositional dominance. In functional data analysis, Wismüller et al. (2000) presented a neural network approach to hierarchical unsupervised clustering of functional magnetic resonance imaging (fMRI) time-sequences of the human brain by self-organized fuzzy minimal free energy vector quantization. Shimizu (2011) proposed a nonparametric method which partitions the curves into clusters and discretizes the dimensions of the curve points into intervals. The cross-product of these partitions forms a data-grid which is obtained using a Bayesian model selection approach while making no assumptions regarding the curves. Boulé et al. (2014) proposed hierarchical clustering for interval-valued functional data as the extension of functional clustering method, and applied this method to real data.

Concerning clustering algorithms for mixed feature-type data, Ichino and Yaguchi (1994) presented simple and convenient generalized Minkowski metrics on the multidimensional feature space in which coordinate axes are associated with not only quantitative features but also qualitative and structural features. Then an approach to the hierarchical conceptual clustering for mixed feature data was proposed. De Carvalho and De Souza (2010) presented an unsupervised pattern recognition methods for mixed feature-type symbolic data based on dynamical clustering methodology with adaptive distances.

However, none of these former methods are able to manage mixed feature-type complex data. In this paper, we present dissimilarity matrix on the multidimensional mixed feature-type complex space, and propose agglomerative hierarchical clustering method for mixed feature-type complex data based on the dissimilarity matrix. The remainder of the paper is organized as follows. Section 2 first describes mixed feature-type complex data, and then introduces dissimilarity matrix on the multidimensional mixed feature-type complex space. Hierarchical clustering method for mixed feature-type complex data is presented using the dissimilarity matrix. To show the usefulness of the method, experiments with simulated data in a framework of a Monte Carlo schema are considered in Section 3. Finally, Section 4 gives the concluding remarks.

2. Hierarchical clustering method for mixed feature-type complex data

In this section, we first describe mixed feature-type complex data, then introduce dissimilarity matrices on the multidimensional space, and present hierarchical clustering method for mixed feature-type complex data at last.

Let a generic data table represent the values of $p$ mixed feature-type complex variables $X_1, \ldots, X_p$ on a set $O = \{1, \ldots, n\}$ of $n$ objects. The $i$th object represents a vector of mixed feature-type complex data $x_i = (x_{i1}, \ldots, x_{ip}) (i = 1, \ldots, n)$, which means $x_{ij} = X_j(i)$ can be a single-valued data, an interval symbolic data, a histogram symbolic data, a compositional data or a functional data according to the type of the corresponding variable. Usually when objects are clustered, the dissimilarity between any two objects is indicated by some sort of distance. Define the global dissimilarity of different mixed feature-type complex objects as $D^2 = \{d^2(x_l, x_k)\}, l, k = 1, \ldots, n$, where $d(x_l, x_k)$ measures the dissimilarity between objects $x_l$ and $x_k$.

Define the local dissimilarity as $D_j^2 = \{d^2(x_l^j, x_k^j)\}, l, k = 1, \ldots, n, j = 1, \ldots, p$, where $d(x_l^j, x_k^j)$ measures the dissimilarity between objects $x_l$ and $x_k$ on variable $X_j$.

As the proposed global dissimilarity $D^2$ is determined by sum of dissimilarities $D^2_j$ corresponding to the different variables, we can obtain

$$D^2 = \sum_{j=1}^{p} \xi_j D_j^2,$$  

(1)
where $\xi_j$ is a normalizing factor for the $j$th variable $X_j$. Based on the standardization method of Chavent (1997), we set $\xi_j = \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{p} d^2(x_{i,k}^j, x_{i,k}^j) / (j = 1, \ldots, p)$.

Equation (1) can be also rewritten as $d^2(x_i^j, x_k^j) = \sum_{j=1}^{p} \xi_j d^2(x_{i,k}^j, x_{i,k}^j), l, k = 1, \ldots, n$. $d(x_i^j, x_k^j)$ can be considered as a distance if the main properties that define a distance are achieved:

1. Reflexivity: $d(x_i^j, x_i^j) = 0$;
2. Symmetry: $d(x_i^j, x_k^j) = d(x_k^j, x_i^j)$;
3. Triangular inequality: $d(x_i^j, x_k^j) \leq d(x_i^j, x_m^j) + d(x_m^j, x_k^j)$.

According to the description of mixed feature-type complex data, $X_j$ can be different types of variables, which indicates that the measure of local dissimilarity on variable $X_j$ varies depending on the type of variable. Then we will respectively introduce the dissimilarity measures for different types of variables under the properties mentioned above.

A complex variable $X_j$ is a single-valued variable when $X_j(i) = x_i^j \in \mathbb{R}$. In this case, the dissimilarity between pairs of objects can be computed according to the Euclidean distance. The corresponding local dissimilarity is

$$d^2(x_i^j, x_k^j) = (x_i^j - x_k^j)^2, l, k = 1, \ldots, n. \quad (2)$$

$X_j$ is an interval-valued variable when $X_j(i) = x_i^j = [a_i^j, b_i^j] \in [a, b], a, b \in \mathbb{R}, a \leq b$. In this case, the dissimilarity between pairs of objects can be computed according to the Wasserstein distance (Irpino and Verde, 2006). It is recommended that the reader see the original sources for details.

If we suppose an uniform distribution of points, the local dissimilarity between intervals $x_i^j$ and $x_k^j$ is defined as

$$d^2(x_i^j, x_k^j)(m_i^j - m_k^j)^2 + \frac{1}{3} (r_i^j - r_k^j)^2, l, k = 1, \ldots, n, \quad (3)$$

where $m_i^j = \frac{a_i^j + b_i^j}{2}$, $r_i^j = \frac{b_i^j - a_i^j}{2}$.

$X_j$ is a histogram-valued variable when $X_j(i) = x_i^j = \{[a_{ih}^j, b_{ih}^j], p_{ih}^j; [a_{ih+1}, b_{ih+1}^j], p_{ih+1}^j; \ldots; [a_{ihn_j}, b_{ihn_j}], p_{ihn_j}^j\}$, where $[a_{ih}^j, b_{ih}^j]$ is the $h$th interval of $x_i^j$ and $p_{ih}^j$ is associated relative frequency. Let $n_{ij}$ denote the number of subintervals in $x_i^j$. Then, $a_{ih}^j \leq b_{ih}^j$ for all $h=1, \ldots, n_{ij}$, and $\sum_{h=1}^{n_{ij}} p_{ih}^j = 1$.

In this case, the dissimilarity between pairs of objects were computed according to the Wasserstein distance (Irpino and Verde, 2006). It is recommended that the reader see the original sources for details.

$X_j$ is a compositional-valued variable when $X_j(i) = x_i^j = \{x_{i,1}^j, \ldots, x_{i,D}^j\}, x_{i,u}^j > 0, \sum_{u=1}^{D} x_{i,u}^j = 1$. In this case, the dissimilarity between pairs of objects were computed according to the Aitchison distance (Aitchison, 1986), which is related to the clr (centred log-ratios) transformation defined by:

$$clr(x_i^j) = [\ln \frac{x_{i,1}^j}{g(x_i^j)}, \ldots, \ln \frac{x_{i,D}^j}{g(x_i^j)}] = \ln \frac{x_i^j}{g(x_i^j)}, \quad (4)$$

where $g(x_i^j) = (x_{i,1}^j \cdots x_{i,D}^j)^{1/D}$ is the geometric mean of the parts of $x_i^j$.

Note that the distance is invariant under permutation of the parts of a composition. Then the local dissimilarity between compositions $x_i^j$ and $x_k^j$ is defined as $d^2(x_i^j, x_k^j) = \sum_{u=1}^{D} (\log \frac{x_{i,u}^j}{g(x_i^j)} - \log \frac{x_{k,u}^j}{g(x_k^j)})^2, l, k = 1, \ldots, n$.

$X_j$ is a functional-valued variable when $X_j(i) = x_i^j(t), t \in [0, T]$. With functional data, an appropriate dissimilarity measure for two functions measured on some domain is the $L_2$ distance between the two curves: $d^2(x_i^j, x_k^j) = \int_0^T (x_i^j(t) - x_k^j(t))^2 dt$. Assume that the functions are expanded in terms of some basis functions by $x_i^j(t) = \sum_{u=1}^{K} a_{i,u}^j \phi_u(t) = a_i^T \Phi^j(t)$, where $a_i^j$ is the vector of basis coefficients for $x_i^j(t)$, and $\Phi^j(t)$ is the vector of basis functions. Then

$$d^2(x_i^j, x_k^j) = \int_0^T (a_i^j - a_k^j)^T \Phi^j(t) \Phi^j(t)^T (a_i^j - a_k^j) dt$$
$$= \int_0^T (a_i^j - a_k^j)^T W^j (a_i^j - a_k^j)^2 dt, l, k = 1, \ldots, n, \quad (5)$$
where $W^j = \int_0^T \Phi^j(t) \Phi^j(t)^T dt$.

For any orthonormal basis such as the Fourier basis, the Gram matrix $W^j$ is the identity matrix. For other basis functions such as B-Splines, $W^j$ must be calculated by numerical integration.

Based on the local dissimilarity measures for different types of variables introduced above, we can obtain the global dissimilarity measures for pairs of complex mixed feature-type objects using Equation (1).

Once the dissimilarity matrix is calculated, a standard agglomerative or divisive hierarchical clustering procedure can be applied. In this paper, an agglomerative hierarchical clustering method will be proposed for mixed feature-type complex data. We consider the Lance and Williams recurrence relation for Ward’s clustering method (Gordon, 1999).

The specific process of hierarchical clustering method for mixed feature-type complex data is described as follows:

**Step 1** Calculating the local dissimilarity $D^{(2)}_j = [d^2(x_i^j, x_k^j)]$ on variable $X_j (j = 1, \ldots, p)$ using the methods in Section 2.1;

**Step 2** Obtaining the global dissimilarity $D^{(2)} = [d^2(x_i, x_k)]$ using $D^{(2)}_j$ in Step 1 and Equation (1);

**Step 3** Hierarchical clustering for mixed feature-type complex data based on the global dissimilarity $D$ in Step 2;

**Step 4** Evaluating the clustering results.

### 3. Simulation study

To validate the effectiveness of the hierarchical clustering method for mixed feature-type complex data, this section will conduct a Monte Carlo simulation experiment using synthetic data.

To measure the quality of the results furnished by the proposed method, an external validity index, the Marcotorchino modified version of the Rand index (Marcotorchino and El Ayoubi, 1991), and the overall error rate of classification (OERC) (Breiman et al., 1984) are used.

The modified version of the Rand index takes its values from the interval $[0, 1]$, where the value 1 indicates perfect agreement between partitions, whereas values close to 0 correspond to random cluster agreement. The OERC index takes its values in the interval $[0, 1]$ where 0 indicates a perfect agreement between the partitions, unlike the value 1.

Without loss of generality, a complex data set of 5 variables on $n$ objects is created. The data set has 3 classes of equal sizes. $X_1, \ldots, X_5$ are single-valued, interval-valued, histogram-valued, compositional-valued and functional-valued variable, respectively.

Objects in single-valued variable are drawn according to normal distribution $N(\mu_1, \sigma^2_1)$. Objects in interval-valued variable are generated in the same way as Dias and Brito (2013a). Each interval-valued data is organized from $k$ single-valued data, which are drawn according to uniform distribution $U(\delta_1, \delta_2)$.

Objects in histogram-valued variable are also generated in the same way as Dias and Brito (2013b). Each histogram-valued data is organized from $k$ single-valued data, which are drawn according to normal distribution $N(\mu_2, \sigma^2_2)$. The simulation data of compositional-valued variable is generated according to Palarea-Albaladejo et al. (2012). To simulate compositions we first generate vectors $y = [y_1, y_2]$ in $\mathbb{R}^2$ from 2-dimensional normal distributions $N(\mu_3, \Sigma)$. Then those vectors are transformed into compositions in $S^3$ by the inverse transformation of alr transformation: $alr(x^j_i) = [\ln x^j_{1i}, \ldots, \ln x^j_{ki}]$.

For the functional-valued variable, we take the same form as Kato et al. (2012). Objects in Functional-valued variable are simulated by $x^j_i (t) = \sum_{k=1}^{50} \gamma_k Z_k \phi_k(t)$, where $\gamma_k = (-1)^{k+1} k^{-\alpha/2}, Z_k \sim U(\delta_3, \delta_4), \phi_k(t) = \sqrt{2} \cos(k\pi t)$.

The simulation design is described as Tab. 1. We investigate the performance of the model with Monte Carlo simulation studies. In the framework of a Monte Carlo experiment, 500 replications of the previous process were carried out.

The hierarchical clustering method has been performed on these datasets. The 3-cluster partition obtained with the clustering method is compared with the 3-class partition known a priori. The averages and standard deviations of Rand and OERC are calculated.

The values of the average of the Rand index and OERC obtained with the hierarchical clustering method for the synthetic complex data sets are 0.9128 and 0.1167 respectively, while the corresponding standard-deviation are 0.1196 and 0.1558 respectively. The results show that the clustering method is effective.

In order to further analyze the result of experiment, one specific experiment of the 500 simulations with 30 objects is
picked out. In the experiment, the evaluation indexes $Rand = 1$ and $OERC = 0$.
The result in Fig. 1 (a) shows that the dendrogram is well-balanced, which allows us to divide the simulated dataset by three the number of clusters. Furthermore, we conduct a multidimensional scaling (MDS) (Kruskal, 1964) analysis based on the dissimilarity matrix of the simulated dataset picked out.

The result in Fig. 1 (b) shows that the points can be well divided into three clusters, which also indicates the good clustering effect of the hierarchical clustering method for mixed feature-type complex data.

4. Conclusions

To deal with mixed feature-type complex data involves single-valued data, symbolic data, composition data and functional data, the paper constructs dissimilarity matrix on the multidimensional mixed feature-type complex space, and propose agglomerative hierarchical clustering method for mixed feature-type complex data based on the dissimilarity matrix. A simulation study is performed to analyze the effect of the hierarchical clustering method based on the constructed dissimilarity matrix. The results show that the dissimilarity matrix can measure the distance between different mixed feature-type complex observations effectively and the hierarchical clustering method performs well.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China(Grant No. 7142010725) and
the National High Technology Research and Development Program of China (863 Program) (Grant No. SS2014AA012303).

Reference


