Clustering Mixed Data Using Spherical Representaion

Yoshiharu Sato

Hokkaido University, Sapporo, Japan
ysato@main.ist.hokudai.ac.jp

Abstract. When the data is given as mixed data, that is, the attributes take the values in mixture of binary and continuous, a clustering method based on k-means algorithm has been discussed. The binary part is transformed into the directional data (spherical representation) by a weight transformation which is induced from the consideration of the similarity between binary objects and of the natural definition of descriptive measures. At the same time, the spherical representation of the continuous part is given by the use of multidimensional scaling on the sphere. Combining the binary part and continuous part, like the latitude and longitude, we obtained a spherical representation of mixed data. Using the descriptive measures on a sphere, we obtain the clustering algorithm for mixed data based on k-means method. Finally, the performance of this clustering is evaluated by actual data.

1 Introduction

The mixed data is defined such a data that each object is measured by the binary attributes and the continuous attributes simultaneously. Then the each object $\boldsymbol{o}_i$ is denoted by

$$\boldsymbol{o}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i) = (x_{i1}, x_{i2}, \cdots, x_{ip}, y_{i1}, y_{i2}, \cdots, y_{iq}), \quad (i = 1, 2, \ldots, n) \qquad (1)$$

where x_{ir} takes binary value 0 or 1, and y_{it} takes the continuous value.

Recently, the size of data goes on increasing by the development of information technology. Then the feasible clustering method seems to be k-means method or its modifications. In k-means method, the concept of mean and variance of the observed data play the essential role. Then the binary data is transromed into directional data in order to get the natural definition of descriptive measures.

When the mixed data is given, traditional cluster analysis has the essential problem in mixture of the distance between binary data and the distance between continuous data. Then a fundamental idea of this paper is that if we get the spherical representation of the binary data and the continuous data simultaneously, we may combine these two spherical data into one spherical data, that is, one is considered to be a latitude and the other to be a longitude. In order to get the spherical representation of q-dimensional continuous data, we use the concept of multidimensional scaling on q-dimensional sphere so as to keep a distance relation between q-dimensional continuous configuration and q-dimensional spherical configuration.

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part II, LNAI 4252, pp. 94–101, 2006.

2 Transformation of Binary Data Into Directional Data

We assume that the following n binary objects with p attributes are given.

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip}), \quad x_{ia} = 1 \text{ or } 2, \ \ (i = 1, 2, \ldots, n;\ a = 1, 2, \ldots, p)$$

We suppose that each object $\boldsymbol{x}_i$ is weighted by the sum of the value of attributes, i.e. sum of the component of the vector $\boldsymbol{x}_i$. When we denote the weighted vector as $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \cdots, \xi_{ip})$, the components are given by

$$\xi_{ia} = x_{ia} / \sum_{b=1}^{p} x_{ib}, \quad \sum_{a=1}^{p} \xi_{ia} = 1, \quad \xi_{ia} > 0, \tag{2}$$

Then the vectors $\boldsymbol{\xi}_i$ are located on $(p-1)$-dimensional hyperplane in the first quadrant of p-dimensional space. We must introduce a suitable metric function on this hyperplane. Since $\boldsymbol{\xi}_i$ has the property in expression 2, we can use an analogy of a discrete probability distribution, i.e. if we regard $\boldsymbol{\xi}_i$ as a probability, then we are able to introduce Kullback-Leibler divergence as a distance measure, which are defined as follows,

$$D(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \frac{1}{2} \sum_{a=1}^{p} (\xi_{ia} - \xi_{ja}) \log \frac{\xi_{ia}}{\xi_{ja}}$$

When we evaluate Kullback-Leibler divergence between two points, $\boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i + d\boldsymbol{\xi}_i$, up to the second order with respect to $d\boldsymbol{\xi}_i$, the line element in this space is given byt

$$D(\boldsymbol{\xi}_i + d\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) = \frac{1}{2} \sum_{a=1}^{p} d\xi_{ia} \log \frac{\xi_{ia} + d\xi_{ia}}{\xi_{ja}} = \frac{1}{4} \sum_{a=1}^{p} \frac{1}{\xi_{ia}} (d\xi_{ia})^2. \tag{3}$$

This is well known as a chi-square distance. However, since the dimension of this space (hyperplane) is $(p-1)$, we get

$$D(\boldsymbol{\xi}_i + d\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) = \frac{1}{4} \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} \left(\delta_{ab} \frac{1}{\xi_{ia}} + \frac{1}{\xi_{ip}} \right) d\xi_{ia} d\xi_{ib}$$

Then we may consider the hyperplane should be a Riemannian space. The structure of the hyperplane will be discussed by the several geometrical quantities. But we know that the induced metric on a hypersphere in p-dimensional Euclidean is denoted as follows. Using a coordinate $(u_1, u_2, \ldots u_p)$ and $u_1 + u_2 + \cdots + u_p = 1, \quad u_a > 0$, when we denote the hypersphere as follows,

$$\ell_1 = \sqrt{u_1}, \ell_2 = \sqrt{u_2}, \cdots, \ell_{(p-1)} = \sqrt{u_{(p-1)}}, \ell_p = \left\{ 1 - \sum_{b=1}^{p-1} u_b \right\}^{1/2}, \tag{4}$$

the induced metric is given by

$$ds^2 = \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} g_{ab} du_a du_b = \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} \frac{1}{4} \left(\delta_{ab} \frac{1}{u_a} + \frac{1}{u_p} \right) du_a du_b. \tag{5}$$

Then we know that the structure of the hyperplane is a hypersphere.

From this result, we define a directional data, i.e. the data on the unit hypersphere using weighted $\boldsymbol{\xi}_i$ as

$$\ell_{ia} = \sqrt{\xi_{ia}}, \quad (a = 1, \ldots, p)$$

The main advantage using the data on the hypersphere is easy to get a global geodesic distance, because we know the geodesic curve on the hypersphere is the great circle. If we discuss on the hyperplane, we must get the geodesic curve, which is a solution of the geodesic equation.

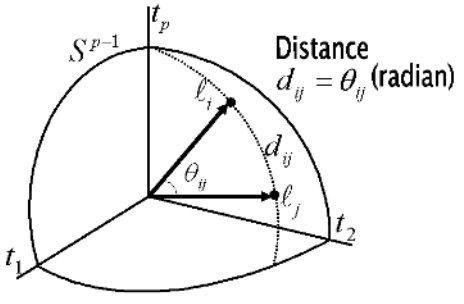

Fig. 1. Directional data and Distance

3 Spherical Representation of Continuous Data

Suppose a q-dimensional continuous data be given by

$$\boldsymbol{y}_i = (y_{i1}, y_{i2} \cdots, y_{iq}), \quad (i = 1, 2, \ldots, n)$$

Using the sample variance and covariance matrix $\boldsymbol{S}$, Mahalanobis distance between a pair of objects $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ is obtained as follows;

$$D = (d_{ij}^2) = (\boldsymbol{y}_i - \boldsymbol{y}_j)\boldsymbol{S}^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j), \quad (i, j = 1, 2, \ldots, n) \tag{6}$$

This distance can be considered as a square of Euclidean distance when the original data $\boldsymbol{y}_i$ is transformed that

$$\boldsymbol{z}_i = \boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}), \quad (i = 1, 2, \ldots, n)$$

where $\bar{\boldsymbol{y}}$ denotes sample mean vector.

We are intended to get the spherical configuration such that the distance between the points on the sphere is consistent with the distance relation between continuous data $\boldsymbol{z}_i$ as much as possible. Then we assign each $\boldsymbol{z}_i$ to a positive quadrant in q-dimensional unit sphere.

q-dimensional unit hypersphere in $(q+1)$-dimensional Euclidean space is denoted as

$$\boldsymbol{x}(\theta_1, \theta_2, \cdots, \theta_q) = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_q \\ x_{q+1} \end{bmatrix} = \begin{bmatrix} \sin\theta_1 \sin\theta_2 \cdots \sin\theta_q \\ \sin\theta_1 \sin\theta_2 \cdots \cos\theta_q \\ \cdots \\ \sin\theta_1 \cos\theta_2 \\ \cos\theta_1 \end{bmatrix}. \tag{7}$$

Let $\boldsymbol{\alpha}$ be a center direction in positive quadrant in q-dimensional unit sphere, and $\boldsymbol{\beta}$ be a center direction perpendicular to the first axis.

$$\boldsymbol{\alpha} = \boldsymbol{x}(\frac{\pi}{4}, \frac{\pi}{4}, \cdots, \frac{\pi}{4}), \quad \boldsymbol{\beta} = \boldsymbol{x}(\frac{\pi}{4}, \cdots, \frac{\pi}{4}, 0). \tag{8}$$

Then, a point on the unit sphere is contained in the positive quadrant if the distance from $\boldsymbol{\alpha}$ is less than

$$\theta^* = \cos^{-1}(\boldsymbol{\alpha}'\boldsymbol{\beta})$$

Hence, the distance relation $D = (d_{ij})$ is transformed as

$$D^* = (d^*_{ij}) = \left\{ \frac{2\theta^*}{d_{\max}} \right\} d_{ij}, \tag{9}$$

where, $d_{\max} = \max_{i,j} d_{ij}$.

We suppose that the data point $\boldsymbol{z}_i$ is assigned to a directional data $\boldsymbol{\ell}_i$. If the distance relation between assigned directional data reproduced the distance relation D^* completely, then

$$\cos d^*_{ij} = \boldsymbol{\ell}_i'\boldsymbol{\ell}_j.$$

When we denote the point on the unit sphere

$$\boldsymbol{\ell}_i(\boldsymbol{\theta}_i) = \begin{bmatrix} \ell_{i1} \\ \ell_{i2} \\ \cdots \\ \ell_{iq} \\ \ell_{i(q+1)} \end{bmatrix} = \begin{bmatrix} \sin\theta_{i1} \sin\theta_{i2} \cdots \sin\theta_{iq} \\ \sin\theta_{i1} \sin\theta_{i2} \cdots \cos\theta_{iq} \\ \cdots \\ \sin\theta_{i1} \cos\theta_{i2} \\ \cos\theta_{i1} \end{bmatrix}, \quad \boldsymbol{\theta}_i = \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ \cdots \\ \theta_{iq} \end{bmatrix}, \tag{10}$$

and we put $Q = (q_{ij}) \equiv \cos d^*_{ij}$, the point $\boldsymbol{\ell}_i$ is obtained so as to minimize

$$\eta = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} (q_{ij} - \boldsymbol{\ell}_i'\boldsymbol{\ell}_j)^2 = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n}, (q_{ij} - \sum_{k=1}^{q+1} \ell_{ik}\ell_{jk})^2 \tag{11}$$

provide that

$$0 \leq \theta_{ik} \leq \frac{\pi}{2},$$

because the point $\boldsymbol{\ell}_i$ lies on the positive quadrant. In order to solve such a optimization problem, we must set an initial values of $\boldsymbol{\theta}_i$, denoted $\boldsymbol{\theta}_i^0$, which are

given as follows; When we denote T_α as the tangent space of the sphere on the point $\boldsymbol{\alpha}$, the dimension of T_α is q and the natural frame is given by

$$\boldsymbol{e}_i = \frac{\partial \boldsymbol{x}}{\partial \theta_i}, \quad (i = 1, 2, \dots q)$$

Normalozing each base $\boldsymbol{e}_i$, we get

$$\boldsymbol{e}_i^* = \frac{\boldsymbol{e}_i}{\| \boldsymbol{e}_i \|}, \quad (1 = 1, \cdots q)$$

By the system $\{\boldsymbol{e}_1^*, \boldsymbol{e}_2^*, \cdots, \boldsymbol{e}_q\}$ and $\boldsymbol{\alpha} = \boldsymbol{e}_{q+1}^*$ is considered to be a orthonormal base of q+1- dimensional Euclidean space. When we denote $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iq})$, the point on the tangent space is described as

$$\boldsymbol{z}_{(T_\alpha)i} = z_{i1}\boldsymbol{e}_1^* + z_{i2}\boldsymbol{e}_2^* + \cdots + z_{iq}\boldsymbol{e}_q^*.$$

Then position vector in $(q+1)$-dimensional Euclidean space is denoted by

$$\boldsymbol{v}_i = \boldsymbol{\alpha} + \boldsymbol{z}_{(T_\alpha)i}.$$

Hence, we put the initial point $\boldsymbol{\ell}_i^0$ as

$$\boldsymbol{\ell}_i^0 = \frac{\boldsymbol{v}_i}{\| \boldsymbol{v}_i \|}.$$

When the mixed data is given by

$$(\boldsymbol{x}_i, \boldsymbol{y}_i) = (x_{i1}, x_{i2}, \cdots, x_{ip}, y_{i1}, y_{i2}, \cdots, y_{iq}),$$

the binary data $\boldsymbol{x}_i$ and the continuous data $\boldsymbol{y}_i$ are represented as the directional data $\boldsymbol{\ell}_i^B(\boldsymbol{\theta}_i^B)$ and $\boldsymbol{\ell}_i^C(\boldsymbol{\theta}_i^C)$, respectively. Then the total $(p+q-1)$-dimensional polar coordinate is given by

$$(\theta_{i1}^B, \cdots, \theta_{i(p-1)}^B, \theta_{i1}^C, \cdots, \theta_{iq}^C) \equiv (\theta_{i1}, \cdots, \theta_{i(p-1)}, \theta_{ip}, \cdots, \theta_{i(p+q-1)}) \equiv \boldsymbol{\theta}_i. \quad (12)$$

Using the polar coordinate $\boldsymbol{\theta}_i$, we get the spherical representation of mixed data, that is, the transformation the mixed data into directional data as follows;

$$\boldsymbol{\ell}_i(\boldsymbol{\theta}_i) = \begin{bmatrix} \sin\theta_{i1} \cdots \sin\theta_{ip} \sin\theta_{i(p+1)} \cdots \sin\theta_{i(p+q-1)} \\ \sin\theta_{i1} \cdots \sin\theta_{ip} \sin\theta_{i(p+1)} \cdots \cos\theta_{i(p+q-1)} \\ \cdots \\ \sin\theta_{i1} \cos\theta_{i2} \\ \cos\theta_{i1} \end{bmatrix}. \quad (13)$$

4 *k*-Means Method for Directional Data

The descriptive measures in directional data are given as follows. We suppose that the directional data on S^{p-1} with the size n is given by

$$\boldsymbol{\ell}_i = (\ell_{i1}, \ell_{i2}, \cdots, \ell_{ip}), \quad \boldsymbol{\ell}_i'\boldsymbol{\ell}_i = 1, \quad (i = 1, 2, \dots, n). \quad (14)$$

The mean direction is given by ([2])

$$\bar{\boldsymbol{\ell}} = (\bar{\ell}_1, \cdots, \bar{\ell}_p), \quad \bar{\ell}_a = \sum_{i=1}^{n} \ell_{ia} \Big/ R, \quad R^2 = \sum_{b=1}^{p} \left(\sum_{j=1}^{n} \ell_{jb} \right)^2 .$$

The variance, called circular variance around mean is known ([2]) as

$$V = \frac{1}{n} \sum_{i=1}^{n} \left\{ 1 - \boldsymbol{\ell}_i' \bar{\boldsymbol{\ell}} \right\} .$$

By the natural extension the k-means algorithm to the directional data, we get the following algorithm (spherical k-means, in short). We suppose that a set of n directional objects on the hypersphere S^{p-1} is given by (14). When the number of clusters K is given, the criterion of spherical k-means algorithm is given by

$$\eta = \sum_{k=1}^{K} \sum_{\ell_i \in C_k} V^{(k)} = \sum_{k=1}^{K} \sum_{\ell_i \in C_k} \left\{ 1 - \boldsymbol{\ell}_i' \bar{\boldsymbol{\ell}}^{(k)} \right\},$$

$$\bar{\boldsymbol{\ell}}^{(k)} = (\bar{\ell}_1^{(k)}, \cdots, \bar{\ell}_p^{(k)}), \quad \bar{\ell}_a^{(k)} = \sum_{\ell_i \in C_k} \ell_{ia} \Big/ \sqrt{\sum_{a=1}^{p} \left(\sum_{\ell_i \in C_k} \ell_{ia} \right)^2} .$$

Minimization η is attained by the maximization of the term $\boldsymbol{\ell}_i' \bar{\boldsymbol{\ell}}^{(k)}$. This term denotes the cosine of the angle between $\boldsymbol{\ell}_i$ and $\bar{\boldsymbol{\ell}}^{(k)}$, that is, the distance between the points on the hypersphere $\boldsymbol{\ell}_i$ and $\bar{\boldsymbol{\ell}}^{(k)}$. Therefore, each point $\boldsymbol{\ell}_i$ is assigned the cluster which has the nearest to its mean.

5 The Performance and Characteristic Feature of the Spherical Clustering

Here we discuss the characteristic fearture of the spherical k-means, proposed here, using actual data set.

First example is a credit card approval data which is submitted by Quinlan, J. R. ([3]) to "The Machine Learning Database Repository". This dataset is interesting because there is a good mixture of attributes. We use 10 binary attributes and 6 continuous attributes. All attribute names and values have been changed to meaningless to protect confidentiality of the data. There two classes in this data, one is approved class the other is not approved class, these denoted "+" and "-". Number of observation of each class is 285 and 356, respectively. We transform this data into directional data, and applied the spherical k-means clustering. In k-means algorithm, we must set the initial seed points (initial class centers). Here we use two different observations which are select from the total observations randomly as the initial seed points. Since k-means algorithm could not guarantee the global optimum solution, this processes are repeated $10,000 \times 15$ times in order to get the local solutions. The result in Table 1.(a) has the minimum within variance in this experiment. Since this data is well-known, there

Table 1. Credit card data(KM: Spherical k-Means Method)

		Observed	
		+	−
KM	+	285	27
	−	0	329
Total		285	356

WV: 0.00887, Ac:95%

(a)

		Observed	
		+	−
KM	+	281	0
	−	4	356
Total		285	356

WV: 0.00891, Ac:99%

(b)

WV : Within Variance, Ac : Accuracy

are many reports on the result of discriminant analysis. But the accuracies are not so good. For the reference, in Table 2 (a), (b), (c), the results of discriminant analysis are shown. Table 1, (b) shows that these two classes are almost linear separable on the sphere. It will be understood that discriminant analysis and cluster analysis are the different criterion. Then the within variances of the discrimination are greater than the result of k-means. Moreover, the criterion of the discrimination is minimize a risk function, usually, the misdiscrimination rate, then the data is processed under the labeled data. But clustering does not take into account the label of the data. However, this result suggest that the classical prototype discrimination method seems to be useful when the data has some structure, that is, gather in clusters. And also we will obtain clusters for mixed data in a natural way by the spherical representaion.

Moreover, the spherical k-means method is essentially the same with ordinal k-means method. Then this property does not depend on the spherical representation. In order make sure that, we apply k-means to ordinal continuous data.The data is Wisconsin Diagnostic Brest Cancer.([3]) This has 30 continuous attributes, namely the usual multivariate data. Total observations are 569. There are two classes, one is malignant cancer, 212 observations, and the other is benign cancer, 357 observations. The result of k-means method and discriminant analysis are shown in Table 3,(a), Table 4. The result of k-means method is almost the same with support vector machine. Most interesting point is Table 3, (b). This shows that this data is completely linear separable. However, this

Table 2. Credit card data(Discriminant Functions)

		Observed	
		+	−
SD	+	244	67
	−	41	289
Total		285	356

WV: 0.00986, Ac:83.2%

(a)

		Observed	
		+	−
LDF	+	253	90
	−	32	266
Total		285	356

WV: 0.120, Ac:81.0%

(b)

		Observed	
		+	−
SVM	+	269	55
	−	16	301
Total		285	356

WV: 0.0101, Ac:88.9%

(c)

SD : Bayse Discriminant function using Spherical Distribution.
LDF : Linear Discriminant Function, SVM : Support Vector Machine.

Table 3. Wisconsin Diagnostic Brest Cancer (KM: k-Means Method)

		Observed	
		M	B
KM	M	200	0
	B	12	357
Total		212	357

WV: 28.63, Ac:98%

(a)

		Observed	
		M	B
KM	M	212	0
	B	0	357
Total		212	357

WV: 29.49, Ac:100%

(b)

M : Malignant Cancer B : Benigh Cancer

Table 4. Wisconsin Diagnostic Brest Cancer (Discriminant Functions)

		Observed	
		M	B
LDF	M	194	2
	B	18	355
Total		212	357

WV: 29.88, Ac:96.5%

(a)

		Observed	
		M	B
SVM	M	205	0
	B	7	355
Total		212	357

WV: 29.94, Ac:98.8%

(b)

LDF : Linear Discriminant Function, SVM : Support Vector Machine

solution is not the solution of k-means method but also the hyper plane which is the perpendicular bisector between means of two classes is not LDF function. It is natural that these classes are linear separable when we observed the attributes which are closely related to the discrimination.

References

1. MacQueen, J. : Some methods for classification and analysis of multivariate observations. In L.M.Le Can & J. Neyman (Eds), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, **1** (1967) 281–297
2. Mardia, K. : Statistics of Directional Data, Academic Press, (1972)
3. UCI Machine Learning Information / The Machine Learning Database Repository (http://www.ics.uci.edu/ mlearn/)