INVITED TALK: TA03-3

Questionnaire Data Analysis Based on a New Dendrogram for Visualization of Hierarchical Structure of Proximity

Speaker: Prof. Takeshi Furuhash (Nagoya University, Japan)

Abstract: Visualization is one of the most useful methods to understand similarity or dissimilarity among data. However, most linear methods such as PCA or MDA are not appropriate for construction of hierarchical model based on the distances (similarity/dissimilarity) between instances, because they do not consider the distances for visualization. This paper proposes a new dendrogram map based on multi-dimensional scaling (MDS) using a spring model which considers the distances between instances. This paper applies the proposed method to clustering of two sets of data. One is those prepared for a thought experiment. The other is questionnaire data on a planned product. The results show that the proposed dendrogram visualizes differences in distances between data while the conventional dendrogram and the conventional MDS were not able to do.

Biography:

- 1985 PhD, Department of Electrical & Electronics Engineering, Nagoya University, Japan
- 1985-1988 Engineer, Toshiba Corporation
- 1988-1990 Assistant Professor, School of Engineering, Nagoya University, Japan
- 1990-2000 Associate Professor, School of Engineering, Nagoya University, Japan,
- 2001-2004 Professor, School of Engineering, Mie University, Japan
- 2004- Professor, Graduate School of Engineering, Nagoya University

Questionnaire Data Analysis Based on a New Dendrogram for Visualization of Hierarchical Structure of Proximity

Minh Tuan Pham*, Tomohiro Yoshikawa*, Takeshi Furuhashi*,
*Nagoya University

Abstract— **Visualization is one of the most useful methods to understand similarity or dissimilarity among data. However, most linear methods such as PCA or MDA are not appropriate for construction of hierarchical model based on the distances (similarity/dissimilarity) between instances, because they do not consider the distances for visualization. This paper proposes a new dendrogram map based on multi-dimensional scaling (MDS) using a spring model which considers the distances between instances. This paper applies the proposed method to clustering of two sets of data. One is those prepared for a thought experiment. The other is questionnaire data on a planned product. The results show that the proposed dendrogram visualizes differences in distances between data while the conventional dendrogram and the conventional MDS were not able to do.**

I. Introduction

Visualization is one of the most useful methods to understand similarity or dissimilarity among high-dimensional data. It is a popular approach to use linear methods such as principal component analysis (PCA) or multiple discriminant analysis (MDA) to visualize the data distribution. PCA calculates an eigen-value decomposition of a data covariance matrix. MDA finds a linear combination of features which best separates two or more classes of objects. But, because these linear methods do not consider distances between instances for the visualization, they are not appropriate for the construction of distance-based hierarchical model. For example, when the data of each class follows a mixture of Gaussians distribution, there could be cases that some classes are not separated in a visible space. Another conventional visualization methods is dendrogram which is a tree diagram frequently used to illustrate the nearest clusters produced by hierarchical clustering. The dendrogram is useful to construct a hierarchical model, but unable to visualize the data distribution. Furthermore, the dendrogram becomes too complex to grasp the hierarchical structure when the number of data becomes large. It is able to visualize a hierarchical model by plotting data on the first two principal coordinate axis with spanning tree[1]. But, because the distances between two data plotted on the first two principal coordinate axes are not considered, the distances in the original space are not kept in the visualized space. For example, the two nearest data in the original space can be plotted farthest in the visible space.

This paper proposes a new dendrogram map based on multi-dimensional scaling (MDS)[2], [3], [4] using a spring model which considers distances between instances. This paper applies the proposed method to clustering of two sets of data. One is those prepared for a thought experiment. The other is questionnaire data on a planned product. The results show that the proposed dendrogram visualizes differences in distances between data while the conventional dendrogram and the conventional MDS were not able to do.

II. Proposed Method

This section describes the method to construct the dendrogram map based on new multi-dimensional scaling method using a spring model which considers distances between instances.

A. Multi-Dimensional Scaling Method using Spring Model

The proposed method keeps the distances between instances in the visualized space to be the same as those in the original space. It is obtained by Multi-Dimensional Scaling (MDS) with a spring model. This paper defines a distance between two instances $i, j \in 1, \ldots, n$ as $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|; i, j \in \{1, \ldots, n\}$, where $\xi = \{\mathbf{x}_l \in \mathbf{R}^m, l = 1, \ldots, n\}$ is a set of m dimensional vectors. In the visual space, the data coordinate $\chi = \{\boldsymbol{x}_l \in \mathbf{R}^2, l = 1, \ldots, n\}$ can be calculated by minimizing the energy

$$E(\chi) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{i,j} \left(d^*_{i,j} - d_{i,j}\right)^2, \tag{1}$$

where, $d^*_{i,j} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|; i, j \in \{1, \ldots, n\}$ is the distance between two instances in visual space. $k_{i,j}$ is a control parameter, which is called the spring coefficient. The distance between the two instances in the visible space becomes equal to the distance in the original space when $k_{i,j}$ becomes larger. Then, this paper proposes the method to minimize the energy $E(\chi)$ by solving a problem of balance of the spring model as follows:

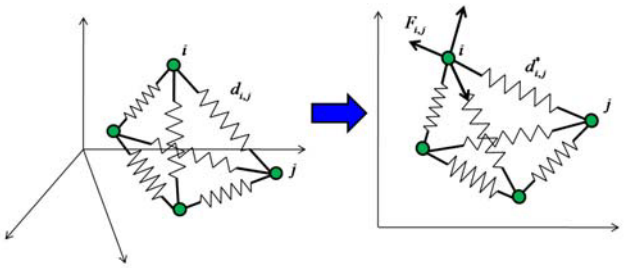

Fig. 1. Multi-dimensional scaling method using spring model

Fig. 1 shows the multi-dimensional scaling method using spring model. The left shows the spring model in the original space. For all pairs from all instances, it makes the link between instance i and instance j by the spring which has a natural length $d_{i,j}$ and a spring coefficient $k_{i,j}$. The spring model is transformed when it is mapped into the visual space in the right figure. So, the potential energy of the

spring model is equivalent to the energy $E(\chi)$ in eq. 1. When all instances were balanced by springs, the potential energy became the smallest. It means that, it can minimize the energy $E(\chi)$ in eq. 1 by solving the problem of balance of the springs of all pairs.

The right in Fig. 1 shows the placement of the instance and the forces F_{ij} acting by the springs in the visible space. The total of the forces acting by all springs to instance i is calculated as,

$$F(i) = \sum_{j=1;j\neq i}^{n} k_{i,j}\left(d_{i,j} - d^*_{i,j}\right)\frac{\boldsymbol{x}_j - \boldsymbol{x}_i}{d^*_{i,j}}. \quad (2)$$

When instance i is balanced by all other instances, $F(i) = 0$. The following formula is obtained:

$$\sum_{j=1}^{n} k_{i,j}\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right) = \sum_{j=1}^{n} k_{i,j}\frac{d_{i,j}}{d^*_{i,j}}\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right). \quad (3)$$

Therefore, about all $i \in \{1, \ldots, n\}$,

$$k_{i,1}\boldsymbol{x}_1 + \ldots - \sum_{j=1}^{n} k_{i,j}\boldsymbol{x}_i + \ldots + k_{i,n}\boldsymbol{x}_n = \boldsymbol{f}_i,$$

$$\boldsymbol{f}_i = \sum_{j=1}^{n} k_{i,j}\frac{d_{i,j}}{d^*_{i,j}}\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right) \in R^2, \quad (4)$$

where, $\boldsymbol{f}_i$ is a 2 dimensional vector. And, the following simultaneous equations are obtained:

$$\begin{aligned}
\mathbf{K} &= \begin{bmatrix} -\sum_{j\neq i}^{n} k_{1,j} & \cdots & k_{1,n} \\ \vdots & \ddots & \vdots \\ k_{n,1} & \cdots & -\sum_{j\neq i}^{n} k_{n,j} \end{bmatrix}, \\
\mathbf{X} &= \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}, \mathbf{F} = \begin{bmatrix} \boldsymbol{f}_1^T \\ \vdots \\ \boldsymbol{f}_2^T \end{bmatrix}, \\
\mathbf{KX} &= \mathbf{F}. \quad (5)
\end{aligned}$$

The coordinates of all instances in visible space are calculated by solving simultaneous equation (5) based on Successive Over-Relaxation(SOR)[5], [6] method. This method is an iterative technique that solves the left hand side of this expression for $\mathbf{X}$, using previous value for $\mathbf{X}$ on the right hand side. Analytically, this is written as:

$$\begin{aligned}
\epsilon^{(m+1)} &= \mathbf{D}^{-1}\left(\mathbf{F} - \mathbf{E}\mathbf{X}^{(m)}\right), \\
\mathbf{X}^{(m+1)} &= \mathbf{X}^{(m)} + \gamma\left(\epsilon^{(m+1)} - \mathbf{X}^{(m)}\right), \quad (6)
\end{aligned}$$

where, $\mathbf{D}$ is a diagonal component of $\mathbf{K}$, and $\mathbf{E} = \mathbf{K} - \mathbf{D}$ is the sum of strictly lower and upper triangular components of $\mathbf{K}$. γ is a convergence speed parameter of the simultaneous equation. For symmetric matrix $\mathbf{K}$, this problem can be proven to be convergent provided that $0 \leq \gamma \leq 2$.

B. Two-dimensional Dendrogram Map

As one of the visualization method, the dendrogram is a tree diagram frequently used to visualize the nearest neighbor relationships among the clusters produced by a hierarchical clustering. Hierarchical clustering is initialized as n clusters that contain only one instance each from n instances of data. Then, it calculates the distance $D(C_1, C_2) = \min_{x_i \in C1, x_j \in C2} D(x_i, x_j)$ between two clusters based on the distance $D(x_i, x_j) = d_{i,j}$ between two instances, and merges two nearest clusters. And a hierarchical structure is obtained by repeating this merger till all data are annexed to one cluster.

The dendrogram is useful to visualize the hierarchical structure, but unable to visualize the data distribution. Furthermore, the dendrogram becomes too complex to grasp the hierarchical construction when the number of data is large. This paper proposes a new dendrogram map which considers both the hierarchical structure of data and the data distribution based on the proposed MDS using the spring model. The algorithm of the proposed method is as follows:

- Step1 Connect two near instances by Nearest Neighbor Method [7], and make a hierarchical cluster structure.
- Step2 Set the spring coefficient of MDS between each two instances with a large value for the connected instances and a small value for the others.
- Step3 Draw data and connected links based on the coordinate of data provided by MDS.

III. Experiments and Discussion

A. Preliminary Experiment

This section examined the effectiveness of proposed dendrogram map by using 3-dimensional data in TABLE I. Fig. 2 shows the visualization result of the data in Table I in 3-dimensional space. This section focuses the distance between instance A and instance J. Instance J was far from instance A in comparison to the other instances. Therefore it was expected that instance J was put relatively far from instance A in a visualized space.

TABLE I
Data for Preliminary Experiment

	x	y	z		x	y	z
A	-1.5	0	0	F	1	1	1
B	-1	1	1	G	1	1	-1
C	-1	1	-1	H	1	-1	-1
D	-1	-1	-1	I	1	-1	1
E	-1	-1	1	J	1.5	0	0

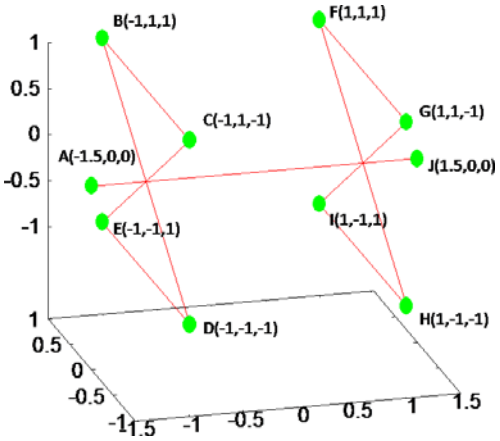

Fig. 2. Data in original space.

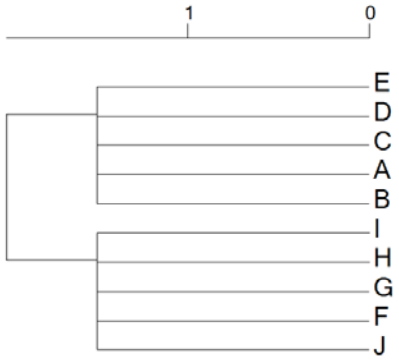

Fig. 3. Dendrogram based on nearest neighbor method.

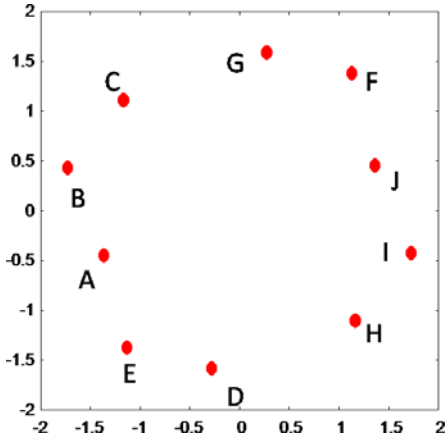

Fig. 4. Data distribution based on conventional MDS.

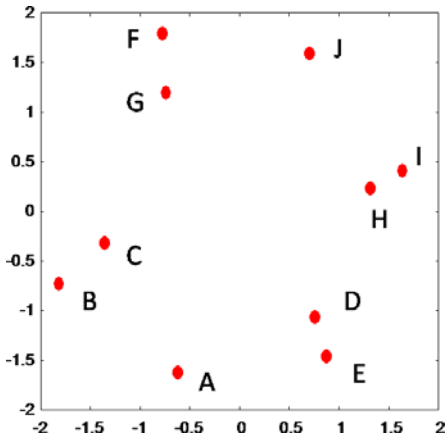

Fig. 5. Data distribution based on proposed method.

Two conventional visualization methods, the conventional dendrogram and the multi-dimensional scaling (MDS), were used to compare their performances with that of the new dendrogram map. Fig. 3 shows the result of visualization using the conventional dendrogram. In Fig. 3, it is found that instance A is near to four instances B, C, D, E; and instance J is near to the other four instances F, G, H, I. Although instance J is located far from instance A, it does not show how far instance J is from instance A. Fig. 4 and Fig. 5 show the results of the conventional MDS and the proposed dendrogram map. The conventional MDS kept the distances between all instances in the original space into the visual space. Therefore, instance J was influenced more by the distances with four instances B, C, D, E, and the result in Fig. 4 shows that instance J was not able to keep the distance with instance A. On the other hand, the proposed dendrogram map puts emphasis on keeping the distances between the nearest data. Instance J was not influenced by the distances with instances B, C, D, E too much, and kept the distance with instance A well. Fig. 5 shows that instance J was put relatively far from instance A in the visualized space.

So, the proposed dendrogram map is expected to be more effective for visualization than the conventional dendrogram and the MDS are.

B. Experimental questionnaire

This experiment involved 1453 respondents. Six scenes that used a new outdoor products α were shown to the respondents for their evaluations. The experiment employed the rating scale method and the respondents were asked to choose one of five grades 1, 2, 3, 4, 5 in response to each of 10 questions per a scene. In this survey, grade 5 means "applicable" while grade 1 means "not applicable". TABLE II shows the 6 scenes (presented by videos during the questionnaire) used as evaluation objects and TABLE III shows the 10 questions.

TABLE II
EVALUATION SUBJECTS.

Object 1	Operating a projector
Object 2	Using a coffee maker and a refrigerator
Object 3	Blogging using PC
Object 4	Using a shower and a dryer
Object 5	Using an electric thruster
Object 6	Using a water purifier

TABLE III
QUESTIONS ASKED FOR POSITIVE IMPRESSIONS.

Q1	It will make me feel superior to those around me.
Q2	Maybe I can perform outdoor activities cleanly.
Q3	It may be useful in emergencies such as disasters
Q4	It may be useful in emergencies such as disasters.
Q5	It looks easy to carry.
Q6	It looks easy to assemble and set up.
Q7	It will make my friends and family happy.
Q8	I will enjoy such activities outdoors.
Q9	Maybe, I cannot enjoy these activities without the product α.
Q10	It will make my outdoor leisure activities more pleasant.

C. Clustering groups of respondents based on visualization results

This section employs interactive clustering in the visualization result of MDS. Before making MDS, this paper defined the feature $\boldsymbol{x}$ of each respondent based on all of the distances between two scenes as follows:

$$\delta_{o_1;o_2} = \sqrt{\sum_{q=1}^{10} (g_{q,o_1} - g_{q,o_2})^2} \quad (7)$$

$$\boldsymbol{x} = \{\delta_{o_1;o_2} \mid \forall o_1, o_2 \in \{1 \ \ldots 6\}\} \quad (8)$$

where, $g_{q,o}$ is the grade of the question q at the scene o. And, $\delta_{o_1;o_2}$ the distances between two scenes o_1 and o_2.

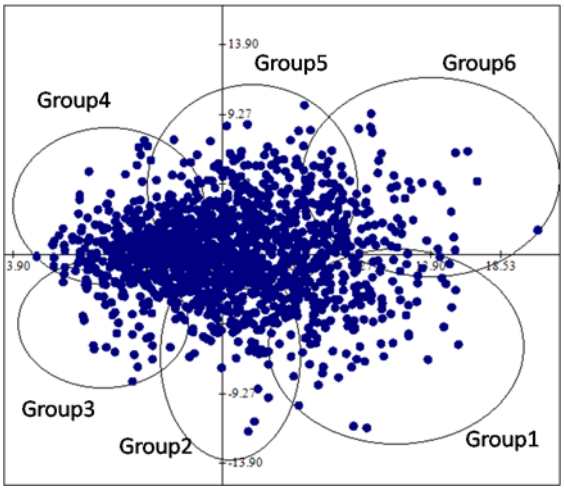

Fig. 6. The visualization result.

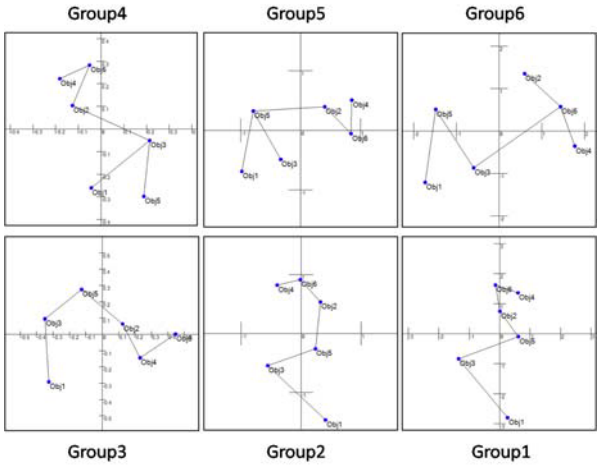

Fig. 8. Obtained dendrogram map by the proposed method.

Fig. 6 shows the visualization result of MDS and the result of interactive clustering. In Fig. 6, six groups were clustered by interactively by the author.

Then, this section applies the proposed dendrogram map and the conventional dendrogram to each group. Fig. 7 shows the conventional dendrogram for six groups in Fig. 6. And Fig. 8 shows the result of proposed dendrogram map.

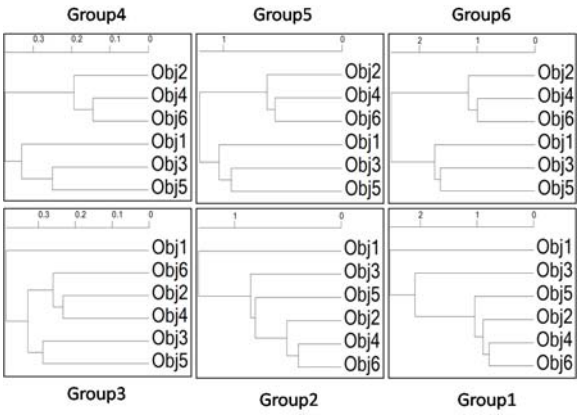

Fig. 7. One-dimensional dendrogram map.

From the dendrograms in Fig. 7, it is found that three objects 2, 4, 6 were considered near by many respondents, and they were far from the other three objects 1, 3, 5. This preference was the same for all six groups in Fig. 7.

From the proposed dendrogram map in Fig. 8, the same preference was also found as in Fig. 7. Furthermore, the proposed method was able to show the differences of these groups. For example, there was a difference of distances between object 2 and object 4 in group 5 and group 6. It means that the respondents in group 5 agreed well to the scenes of object 2 and 4, and those in group 6 do not so much. In the case of group 4, the respondents considered object 5 be the furthest from object 2, 4, 6, although the other groups tend to consider that object 1 was the furthest. The conventional dendrogram was not able to catch the these differences.

IV. Conclusion

This paper proposed the new dendrogram map based on multi-dimensional scaling using a spring model which considers distances between instances. This paper applied the proposed method to clustering of two sets of data. One was those prepared for a thought experiment. This experiment showed that the proposed dendrogram map was able to keep the distances in the visible space better than the conventional dendrogram and the MDS were. The other was questionnaire data on a planned product. The proposed dendrogram map was able to catch the differences between six groups were clustered interactively by the author, which the conventional dendrogram was not. Thus, the proposed dendrogram map has been shown to be a powerful method for visualization data.

References

[1] W. J. Krzanowski, Principles of Multivariate Analysis. A User ' s Perspective. Oxford:Oxford University Press, 1988
[2] W. S. Torgerson, Theory and methods of scaling. New York, Wiley, 1958.
[3] A. Buja, D. F. Swayne, M. Littman, N. Dean, and H. Hofmann. XGvis, Interactive data visualization with multidimensional scaling. Journal of Computational and Graphical Statistics, 2001.
[4] T. Yamada, K.Saito and N.Ueda,Cross-Entropy Directed Embedding of Network Data, Proceedings of the Twentieth International Conference on Machine Learning, p.832–839, 2003.
[5] D. Young, Iterative methods for solving partial difference equations of elliptic type. Trans. Am. Math. Soc. 76, p.92–111, 1954.
[6] C. G. Broyden, Some generalizations of the theory of successive over-relaxation. Num. Math. 6, 269–284, 1964.
[7] B.S.Everitt, Cluster Analysis, Edward Arnold, third edition, 1993.