IEICE TRANS. INF. & SYST., VOL.E95–D, NO.3 MARCH 2012

PAPER

Extrapolation of Group Proximity from Member Relations Using Embedding and Distribution Mapping

Hideaki MISAWA[†a)], ***Student Member***, **Keiichi HORIO**[†,††b)], ***Member***, **Nobuo MOROTOMI**[†††], **Kazumasa FUKUDA**[†††], ***and*** **Hatsumi TANIGUCHI**[†††], ***Nonmembers***

SUMMARY In the present paper, we address the problem of extrapolating group proximities from member relations, which we refer to as the group proximity problem. We assume that a relational dataset consists of several groups and that pairwise relations of all members can be measured. Under these assumptions, the goal is to estimate group proximities from pairwise relations. In order to solve the group proximity problem, we present a method based on embedding and distribution mapping, in which all relational data, which consist of pairwise dissimilarities or dissimilarities between members, are transformed into vectorial data by embedding methods. After this process, the distributions of the groups are obtained. Group proximities are estimated as distances between distributions by distribution mapping methods, which generate a map of distributions. As an example, we apply the proposed method to document and bacterial flora datasets. Finally, we confirm the feasibility of using the proposed method to solve the group proximity problem.

key words: *group proximity, relational data, multidimensional scaling (MDS), self-organizing map (SOM), SOM of SOMs (SOM^2)*

1. Introduction

Data visualization plays an important role in finding proximity structures of hidden information in data. Principal component analysis (PCA) [1] and the self-organizing map (SOM) [2] are representative visualization tools for vectorial data. Most data analysis methods are based on the vectorial representation of data, whereas there exist various types of non-vectorial data, such as sequences, trees, and graphs. For such structured data, pairwise relations between objects are often measured as proximities (similarities or dissimilarities). The data that consist of pairwise proximities are referred to as relational data, proximity data, or (dis)similarity data. Relational data can be created from vectorial data with a certain similarity or dissimilarity measure. Therefore, the relational representation of data may be more general than the vectorial representation [3]. For example, pairwise relational data occur as alignment scores or evolutionary distances between two DNA sequences in bioinformatics and also occur as human proximity judgments in empirical sciences, such as psychology and psychophysics. Multidimensional scaling (MDS) and its variants can visualize proximity structures of relational data [4], [5].

In the present paper, we address the problem of extrapolating group proximities from member relations. We assume that a relational dataset consists of several groups and that pairwise relations of all members can be measured, regardless of the groups to which the members belong. The goal is to estimate group proximities from pairwise relations under this situation. We refer to this problem as the group proximity problem.

Let us present an example of journal similarities. Journals are collections of articles, and each journal is regarded as a group of articles. Articles are considered to be relational data because their pairwise similarities, such as co-occurrence counts of certain words or co-citation counts, can be measured. In this case, a journal and an article are regarded as a group and a member, respectively, and the goal is to estimate journal proximities from article similarities. This example is related to text mining, web mining, and bibliometrics.

As another example, let us consider the case of bacterial flora analysis. Bacterial floras are communities of bacteria, and each bacterial flora is represented as a set of bacterial DNA sequences in gene-based analysis. The properties of bacterial floras are characterized by their compositions (types and relative amounts of bacteria). The proximities of bacterial DNA sequences are calculated as alignment scores or evolutionary distances. In this case, a bacterial flora and a bacterial DNA sequence are regarded as a group and a member, respectively, and the goal is to estimate compositional proximities from sequence similarities. This example is related to (microbial) ecology and is further illustrated in Sect. 3.

In order to solve the group proximity problem, we present a method based on embedding and distribution mapping. In the proposed method, all relational data are transformed into vectorial data by embedding methods. After this process, the vectorial distributions of groups are obtained. Group proximities are estimated as distances between distributions by distribution mapping methods, which generate a map of distributions. In the present paper, we use a metric MDS and the SOM of SOMs (SOM^2) [6] for embedding and distribution mapping, respectively.

The remainder of the present paper is organized as fol-

Manuscript received June 15, 2011.
Manuscript revised October 27, 2011.

[†]The authors are with the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu-shi, 808–0196 Japan.
[††]The author is with the Fuzzy Logic Systems Institute (FLSI), Iizuka-shi, 820–0067 Japan.
[†††]The authors are with the Department of Microbiology, School of Medicine, University of Occupational and Environmental Health, Kitakyushu-shi, 807–8555 Japan.
a) E-mail: misawa-hideaki@edu.brain.kyutech.ac.jp
b) E-mail: horio@brain.kyutech.ac.jp
DOI: 10.1587/transinf.E95.D.804

Copyright © 2012 The Institute of Electronics, Information and Communication Engineers

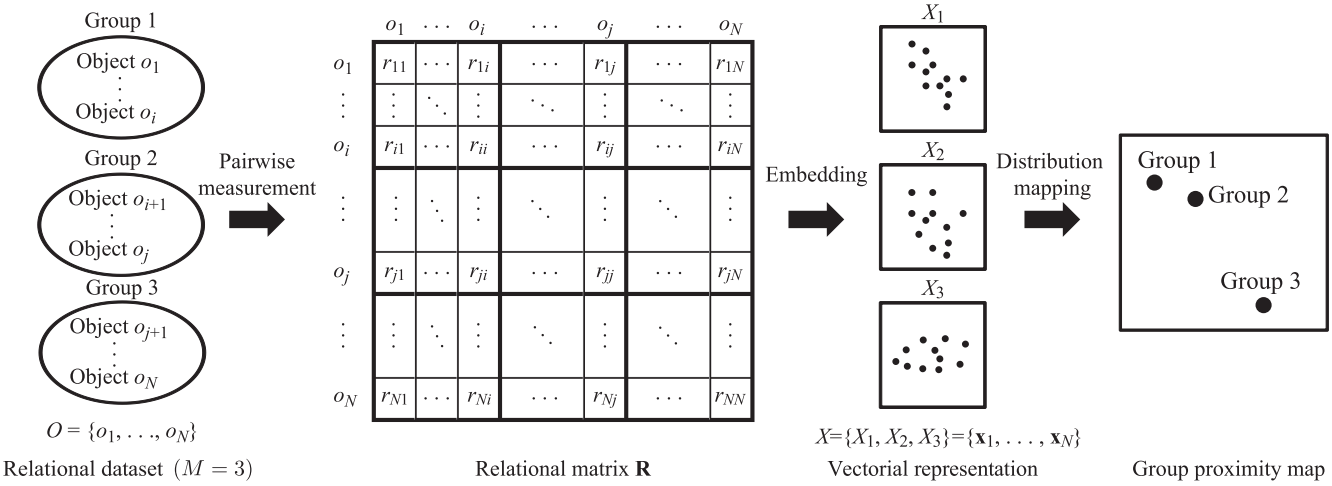


Fig. 1 Process of the proposed method.

lows. In Sect. 2, the problem is formulated, and the process of the proposed method is described. Experimental results obtained for two datasets are presented in Sect. 3. Finally, we conclude the paper in Sect. 4.

2. Group Proximity Problem

2.1 Problem Formulation

The problem considered herein is that of extrapolating group proximities from member relations. Let $O = \{o_1, \ldots, o_N\}$ denote a set of N objects. Since the problem here is not clustering, the objects O are divided into M groups in advance. We assume that the pairwise relations of all objects can be measured regardless of the groups to which the objects belong. Relational data are given in the form of an $N \times N$ relational matrix, $\mathbf{R} = (r_{ij})$, where r_{ij} is the pairwise relation between objects o_i and o_j. A pairwise similarity or dissimilarity is often used as the pairwise relation r_{ij}. In this situation, the problem is to estimate group proximities from the relational matrix $\mathbf{R}$.

Figure 1 shows the process of the proposed method. In the proposed method, the relational data $\mathbf{R}$ are transformed into vectorial data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ by embedding methods. After this process, the vectorial distributions of the groups are obtained. Group proximities are estimated as distances between distributions by distribution mapping methods. The remainder of this section describes the embedding and distribution mapping methods.

2.2 Embedding

Embedding methods represent objects in a low-dimensional Euclidean space in such a way that certain relationships between the objects are preserved. For example, isometric feature mapping (Isomap) [7], local linear embedding (LLE) [8], and stochastic neighbor embedding (SNE) [9] try to preserve geodesic distances, local geometries, and probabilities of objects being neighbors, respectively.

Among such embedding techniques, MDS is one of the most well-known methods for embedding relational data [4], [5]. Metric MDS attempts to preserve pairwise dissimilarities. For the group proximity problem, pairwise relations (dissimilarities) of members should be preserved as distances between embedding vectors because the group proximities are estimated from distribution distances, which are based on vector distances. Hence, a metric MDS is used in the present paper.

Suppose that pairwise relation r_{ij} is given as a pairwise dissimilarity δ_{ij}. The goal of metric MDS is to find a representation of the objects in low-dimensional space so that the distances d_{ij} approximates the dissimilarities δ_{ij}, where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$. In the present paper, we use a metric MDS, in which the cost function, usually referred to as the stress function, is defined as

$$E = \sqrt{\frac{\sum_{i>j}(\delta_{ij} - d_{ij})^2}{\sum_{i>j}\delta_{ij}^2}}. \tag{1}$$

The configuration of $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is optimized by minimizing the cost function E.

2.3 Distribution Mapping

In the present paper, distribution mapping methods are defined as a method for generating a map of distributions, as shown in Fig. 1. If the distances between distributions can be calculated, then the distributions are considered to be relational data. Therefore, embedding methods can be used to generate a map of the distributions. However, it is not easy to calculate distribution distances in analytical form.

The SOM has been used to visualize the similarities between data vectors in a variety of fields [2]. The SOM^2 was proposed as an extension of the SOM so as to represent relationships between data distributions [6]. The SOM^2 generates a map of the distributions. In the learning of the SOM^2, each distribution is represented by a SOM, and distribution distances are estimated using the learned SOMs. The SOM^2

is useful for the group proximity problem because the SOM can represent the different distributions by learning. Hence, the SOM^2 is used in the present paper.

In the following, we briefly explain the architecture and learning process of the SOM^2. For additional details, please refer to [6]. Suppose that M groups of data $\{X_1, \ldots, X_M\}$ are given and each group X_m consists of I_m data vectors $X_m = \{\mathbf{x}_{m1}, \ldots, \mathbf{x}_{mI_m}\}$. A SOM^2 has M child SOMs and a single parent SOM. The m-th child SOM has L reference vectors $\{\mathbf{v}_{m1}, \ldots, \mathbf{v}_{mL}\}$. A joint reference vector $\mathbf{V}_m$ of the m-th child SOM is defined as $\mathbf{V}_m = (\mathbf{v}_{m1}, \ldots, \mathbf{v}_{mL})$. The task of the m-th child SOM is to represent the m-th group distribution of X_m. The parent SOM has K reference maps $\{\mathbf{W}_1, \ldots, \mathbf{W}_K\}$, where the k-th reference map $\mathbf{W}_k$ represents a joint reference vector $\mathbf{W}_k = (\mathbf{w}_{k1}, \ldots, \mathbf{w}_{kL})$. The task of the parent SOM is to generate a SOM of the group distributions using the child SOMs. The learning algorithm of the SOM^2 is described as follows.

Step 0 All reference vectors $\{\mathbf{v}_{ml}(0)\}$ and $\{\mathbf{w}_{kl}(0)\}$ are initialized randomly.

Step 1 Upon iteration of the learning t, the child SOMs are updated by the batch SOM algorithm as follows:

$$l^*_{mi}(t) = \arg\min_l \|\mathbf{x}_{mi} - \mathbf{v}_{ml}(t-1)\|^2, \tag{2}$$

$$\beta^l_{mi}(t) = \frac{\exp\left(-d_c^2(l, l^*_{mi}(t))/2\sigma_c^2(t)\right)}{\sum_{i'=1}^{I_m} \exp\left(-d_c^2(l, l^*_{mi'}(t))/2\sigma_c^2(t)\right)}, \tag{3}$$

$$\mathbf{v}_{ml}(t) = \sum_{i=1}^{I_m} \beta^l_{mi}(t)\mathbf{x}_{mi}, \tag{4}$$

where $l^*_{mi}(\cdot), d_c(\cdot,\cdot), \sigma_c(\cdot), \beta^l_{mi}(\cdot)$ are the index of the best matching unit, the distance between two units on the child SOM grid, the neighborhood size, and the neighborhood coefficient, respectively.

Step 2 The reference maps are updated by the batch SOM algorithm regarding $\{\mathbf{V}_m(t)\}$ as a set of data vectors.

$$\begin{aligned} k^*_m(t) &= \arg\min_k \|\mathbf{W}_k(t) - \mathbf{V}_m(t)\|^2 \\ &= \arg\min_k \sum_{l=1}^{L} \|\mathbf{w}_{kl}(t) - \mathbf{v}_{ml}(t)\|^2, \end{aligned} \tag{5}$$

$$\alpha^k_m(t) = \frac{\exp\left(-d_p^2(k, k^*_m(t))/2\sigma_p^2(t)\right)}{\sum_{m'=1}^{M} \exp\left(-d_p^2(k, k^*_{m'}(t))/2\sigma_p^2(t)\right)}, \tag{6}$$

$$\mathbf{W}_k(t) = \sum_{m=1}^{M} \alpha^k_m(t)\mathbf{V}_m(t). \tag{7}$$

Here, $k^*_m(\cdot), d_p(\cdot,\cdot), \sigma_p(\cdot), \alpha^k_m(\cdot)$ are the index of the best matching map, the distance between two maps on the parent SOM grid, the neighborhood size, and the neighborhood coefficient.

Step 3 The best matching maps are copied to the corresponding child SOMs.

$$\mathbf{V}_m(t) = \mathbf{W}_{k^*_m}(t). \tag{8}$$

Steps 1 through 3 are repeated while reducing the neighborhood sizes $\sigma_p(t)$ and $\sigma_c(t)$ monotonically according to the following equations until t reaches the number of learning iterations t_L:

$$\sigma(t) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \exp(-t/\tau), \tag{9}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum values, respectively, of the neighborhood size.

3. Experiments

In order to illustrate the proposed method, we performed experiments on two datasets. In the first experiment, we used the NIPS 14-16 dataset [†]. Although the NIPS 14-16 dataset is not a genuine relational dataset, we used this dataset as an easily understandable example. In the second experiment, as a genuine relational dataset, we used a bacterial flora dataset, which was a set of bacterial DNA sequences.

3.1 Experiment 1: NIPS Document Dataset

This dataset was created by Globerson [10] from articles presented in the Neural Information Processing Systems (NIPS) conferences from 2001 to 2003. The number of documents in the dataset is 593. Each document was categorized into one of 13 technical areas. In this experiment, a technical area and a document were regarded as a group and a member, respectively. The goal of this experiment is to estimate the proximities between the technical areas from document dissimilarities.

3.1.1 Experimental Setting

According to the experiment in [10], the 100 most frequent words were first removed, and then the next 2,000 most frequent words were used. Each document was represented as a word count vector. The word-count vectors were normalized such that the sum of the word counts in each document was one. One document was excluded because the document was far away from other documents in the MDS coordinates. Thus, a total of 592 documents were used in this experiment. The dissimilarities δ_{ij} between the documents were calculated as Euclidean distances between the normalized word-count vectors, and the 592×592 dissimilarity matrix $\Delta = (\delta_{ij})$ was created.

We used the MDS implemented as the mdscale function in the MATLAB statistics toolbox with cmdscale initialization. In the learning of the SOM^2, the parameters were set as follows: $M = 13$, $L = 25$ (5×5), $K = 36$ (6×6), $\sigma_{p,\max} = 9$, $\sigma_{p,\min} = 1$, $\tau_p = 30$, $\sigma_{c,\max} = 7.5$, $\sigma_{c,\min} = 0.1$, $\tau_c = 40$, and $t_L = 400$.

3.1.2 Results and Discussion

Figure 2 shows a vectorial representation of the NIPS doc-

[†]Available from http://robotics.stanford.edu/~gal/data.html.

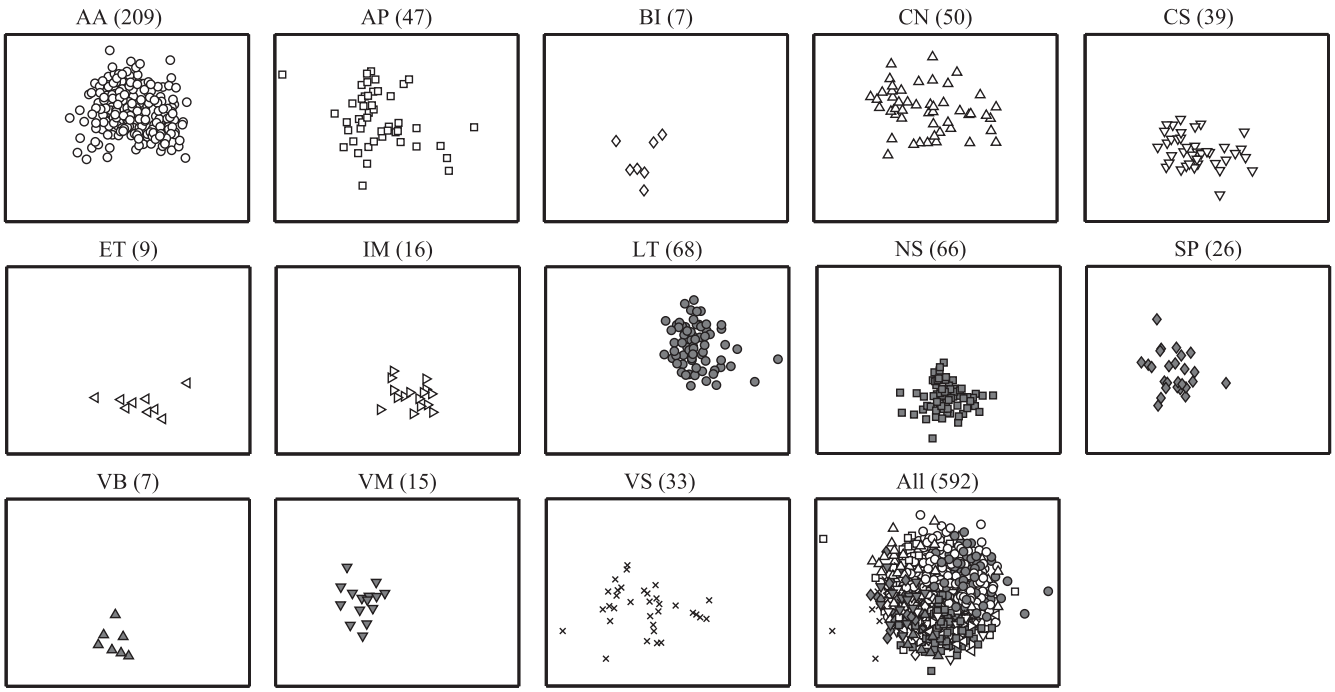


Fig. 2 Vectorial representation of the NIPS dataset by MDS.

uments. All of the documents were transformed into two-dimensional vectorial data by the MDS, as shown in the rightmost bottom panel of Fig. 2. The value of the cost function E was 0.37. Data points have one-to-one correspondences with the documents. Each panel shows the document distribution of each technical area, where the number in parentheses indicates the number of documents in each technical area. Figure 2 shows that the groups (technical areas) tend to be composed of the nearby points (documents) in the MDS coordinates.

If the dissimilarities are the Euclidean distances and the dimensionality of embedding vector is proper, then the MDS can isometrically embed the objects. The dimensionality of the embedding vector should be high in order to preserve the pairwise dissimilarities as much as possible. However, we observed the low reproducibility of the SOM^2 results for the high-dimensional embedding vectors. We believe that the reason for this is that the SOM can approximate the data distribution, the intrinsic dimension of which is high, in several different ways. Therefore, two-dimensional embedding was used for visual interpretation and reproducibility in the present paper.

Figure 3 (a) shows a parent map ($K = 6 \times 6$) generated by the SOM^2 from the distributions shown in Fig. 2. The grid in each box represents the reference map ($L = 5 \times 5$). The distributions shown in Fig. 2 were overlaid on the corresponding best matching maps. Figures 3 (b) and 3 (c) show the U-matrix and MDS representations, respectively, of the parent map. The U-matrix representation visualizes the distances between the neighboring reference maps as color or gray scale and reveals cluster structures [11]. The MDS representation was generated by the MDS from the distances between the reference maps. In both representations, the distances between the reference maps were calculated as the distances between the joint reference vectors in the manner described in Eq. (5).

As shown in Fig. 3, similar technical areas were placed in close proximity to each other, although the numbers of documents in the technical areas were very different. The technical areas related to machine learning, e.g., AA (Algorithms & Architectures), CN (Control & Reinforcement Learning), and LT (Learning Theory), tend to be in the upper right part of the parent map. On the other hand, the technical areas related to brain science, e.g., BI (Brain Imaging) and VB (Biological Vision), tend to be located in the left part of the parent map.

The documents related to vision were divided into three technical areas. Only VS (Vision) was used for the documents in 2001 and 2002, whereas, instead of VS, VB and VM (Machine Vision) were used for the documents in 2003. Documents corresponding to VB and VS are considered to have been included in VS. As a result, VS was located between VB and VM, as shown in Fig. 3.

NS (Neuroscience) and IM (Implementations) were arranged next to each other in Fig. 3 although these technical areas appear not to be similar. The reason for this is that the

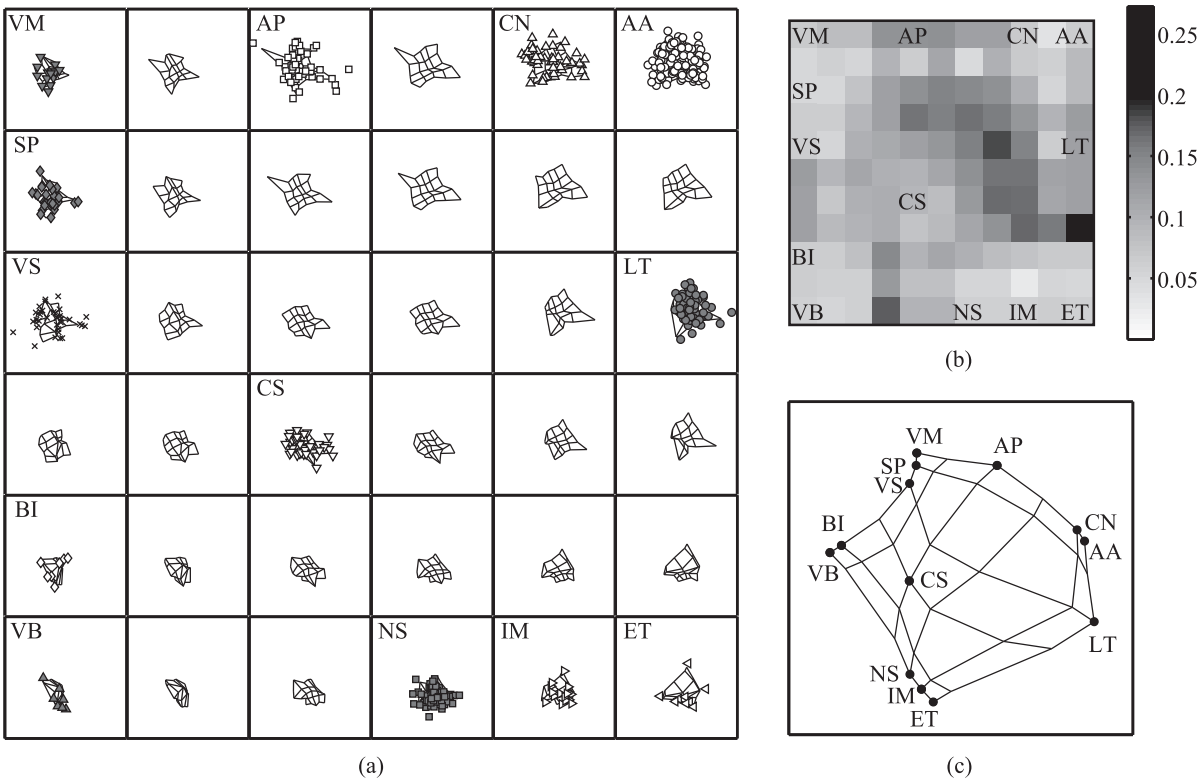


Fig. 3 Group proximity map for the NIPS dataset. (a) Parent map of the SOM^2. (b) U-matrix representation of the parent map. (c) MDS representation of the parent map.

documents in IM were concerned primarily with the hardware implementation of models related to NS, e.g., neurons and neural networks.

We demonstrated that the proposed method can visualize the proximities between the technical areas from document dissimilarities. In this experiment, group proximities largely depend on the positions of the groups in the MDS coordinates. In the next experiment, we present an example in which group proximities depend on their compositions.

3.2 Experiment 2: Bacterial Flora Dataset

We applied the method to a bacterial flora dataset in the second experiment. Bacterial floras are communities of bacteria and their properties are characterized by their compositions, which are the types and relative amounts of bacteria contained in the floras. The types (taxa) of bacteria are identified by 16S ribosomal RNA (rRNA) gene analysis, where bacterial taxa are estimated based on the similarities of 16S rRNA gene sequences to the sequences of known species. In bacterial flora analysis, it is necessary to elucidate relationships between bacterial compositions and physiological or environmental conditions. In the first stage of the analysis, questions such as whether the compositions of bacterial floras are similar and what part of a bacterial flora is different from other bacterial floras arise. In this experiment, we attempted to answer these questions using the proposed method.

In this experiment, we analyzed a part of sequence data used in [12]. Morotomi et al. analyzed the intestinal bacterial floras of 29 healthy Japanese adults [12]. The fecal samples were collected from 29 subjects twice at five month intervals. In order to identify bacteria contained in each sample, 96 clones were randomly selected and sequenced from each sample using a clone library method [13]. In this experiment, for the sake of easy interpretation, 10 samples (A through J) were selected from among 29 samples in the first collection (samples 1, 6, 8, 11, 16, 17, 24, 26, 28, and 29 in [12]). After refining sequences, we obtained 887 sequences of the partial 16S rRNA genes in total. In this experiment, a bacterial flora (sample) and a sequence were regarded as a group and a member, respectively. The goal of this experiment is to estimate proximities between bacterial floras from sequence dissimilarities.

3.2.1 Experimental Setting

The similarity between two DNA sequences can be measured as an alignment score. The alignment is the process of lining up two sequences to assess their degree of similarity. We used the Smith-Waterman (SW) alignment algorithm [14], which was implemented as the SSEARCH program in the FASTA sequence analysis package [15]. Let s_{ij} denote the SW score between the i-th sequence and the j-th sequence. The SW score s_{ij} was linearly transformed into the dissimilarity δ_{ij} as follows:

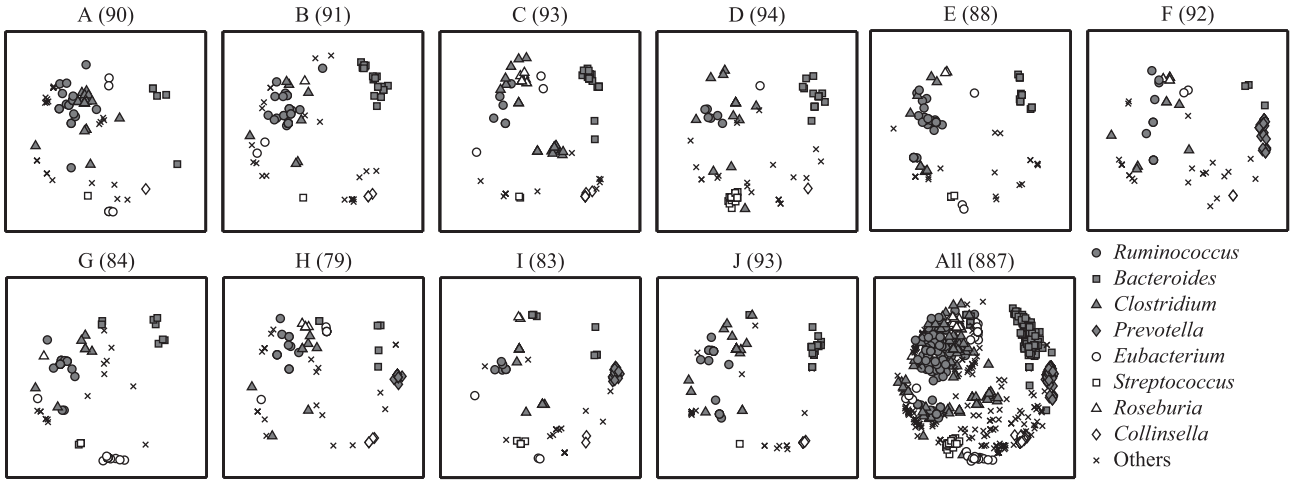


Fig. 4 Vectorial representation of the bacterial flora dataset by MDS.

$$\delta_{ij} = \begin{cases} 1 - s_{ij}/s_{\max} & (i \neq j) \\ 0 & (i = j) \end{cases}, \tag{10}$$

where $s_{\max}$ is the maximum SW score. The 887×887 dissimilarity matrix $\Delta = (\delta_{ij})$ was used in the MDS. We used the MDS implemented as the mdscale function in the MATLAB statistics toolbox with cmdscale initialization. In the learning of the SOM^2, the parameters were set as follows: $M = 10$, $L = 49$ (7×7), $K = 36$ (6×6), $\sigma_{p,\max} = 9$, $\sigma_{p,\min} = 1$, $\tau_p = 30$, $\sigma_{c,\max} = 10.5$, $\sigma_{c,\min} = 0.1$, $\tau_c = 40$, and $t_L = 400$.

3.2.2 Results and Discussion

Figure 4 shows a vectorial representation of the bacterial sequence data. All bacterial sequences were transformed into two-dimensional vectorial data by the MDS, as shown in the rightmost bottom panel of Fig. 4. The value of the cost function E was 0.35. Data points have one-to-one correspondences with bacterial DNA sequences. Different symbols indicate different bacterial taxa at the taxonomic rank of genus. A total of 46 genera were detected from 10 samples. The eight most frequent genera are shown by various different symbols, and the remaining genera are shown as "others". The number in parentheses in each panel denotes the number of sequences in each sample. The bacterial flora of each sample was represented as the vectorial distribution. Unlike the NIPS dataset (Fig. 2), the groups (bacterial flora) did not consist of the nearby points (sequences) in the MDS coordinates. Figure 5 shows the composition of each sample. Figures 4 and 5 show that the samples had different compositions.

Since the dissimilarities used in this experiment were non-Euclidean distances, the MDS did not isometrically embed the objects. Furthermore, two-dimensional embedding was used for visual interpretation and reproducibility, which led to the low preservation of the pairwise dissimilarities. The effects of the non-Euclideanity and preservation accuracy on the results must be investigated further.

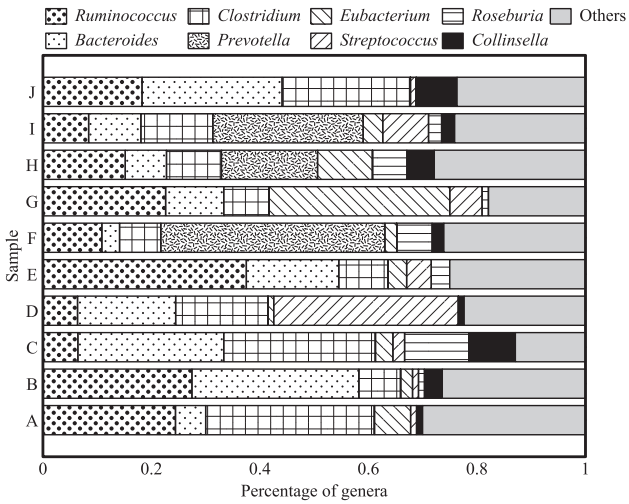


Fig. 5 Composition of each sample.

Figure 6 (a) shows a parent map ($K = 6 \times 6$) generated by the SOM^2 from the data distributions shown in Fig. 4. The grid in each box represents the reference map ($L = 7 \times 7$). The data distributions shown in Fig. 4 were overlaid on the corresponding best matching maps. Note that the information of bacterial taxa was not used in the SOM^2 algorithm. Figures 6 (b) and 6 (c) show the U-matrix and MDS representations, respectively, of the parent map.

Figure 6 shows that the distributions of samples I and F are more similar than those of the other samples. In the diagonal direction from lower right to upper left on the parent map, the percentages of the genus *Ruminococcus* and the genus *Prevotella* tend to decrease and increase, respectively. This direction corresponds to the horizontal axis in Fig. 6 (c). On the other hand, in the diagonal direction from lower left to upper right on the parent map, the percentages of the genus *Streptococcus* and the genus *Collinsella* tend to decrease and increase, respectively. This direction corresponds to the vertical axis in Fig. 6 (c). It is confirmed that the proposed method can reveal the compositional dif-

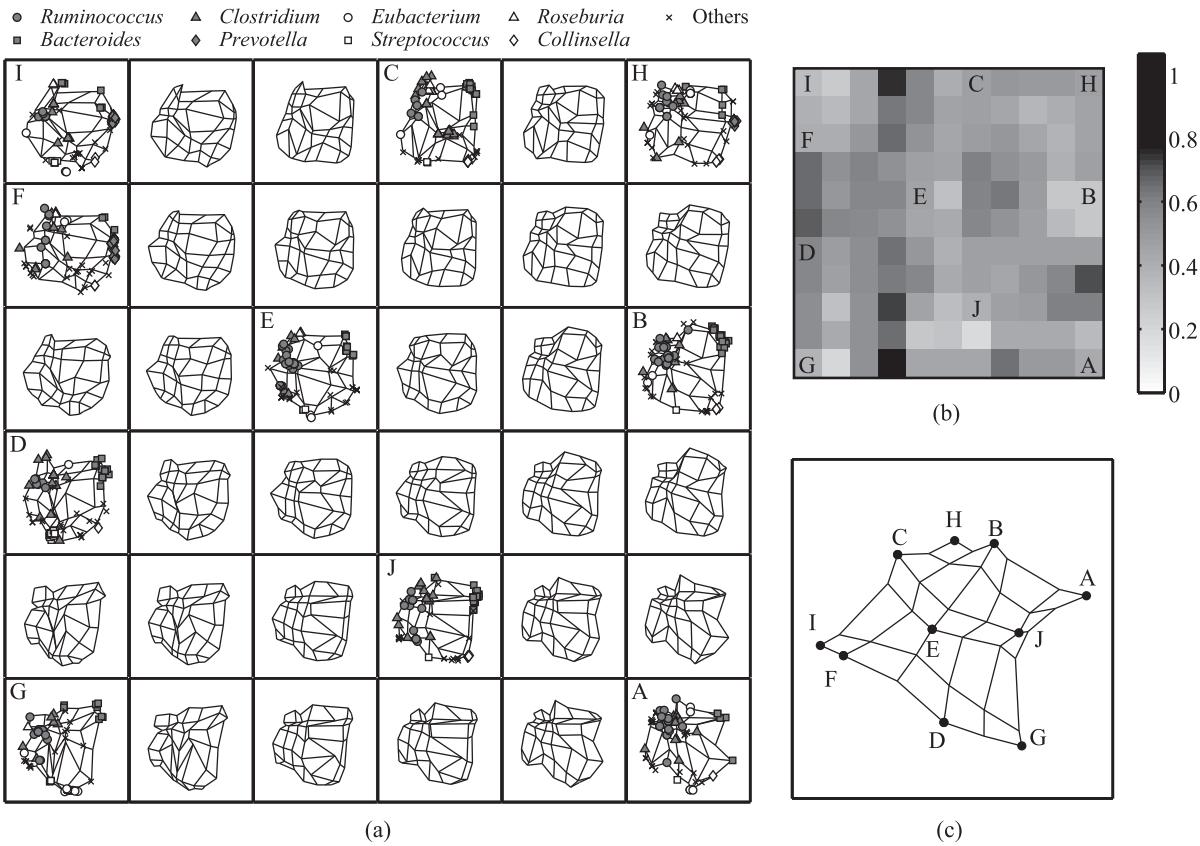


Fig. 6 Group proximity map for the bacterial flora dataset. (a) Parent map of the SOM^2. (b) U-matrix representation of the parent map. (c) MDS representation of the parent map.

ferences.

As shown in Fig. 5, the genus *Streptococcus* and the genus *Eubacterium* were dominant in samples D and G, respectively. Samples D and G appear to differ in their composition. However, Fig. 6 indicates that these samples are similar. This is because the sequences of the genus *Streptococcus* in sample D are similar to the sequences of the genus *Eubacterium* in sample G, as shown in Fig. 6. This result implies that the proposed method can take into account the sequence (member) similarities. The similarity between sequences of 16S rRNA genes is used to identify bacterial taxa. Therefore, this sequence similarity is thought to implicitly represent the similarity between properties of bacteria. The proposed method can be used for bacterial flora analysis, in which the sequence similarity must be taken into account.

In this experiment, the compositional proximities of the bacterial flora were estimated based on the distributions. The next stage of bacterial flora analysis is to reveal the relationships between bacterial compositions and physiological or environmental conditions. The proposed method may contribute to this stage of analysis.

Finally, we emphasize the potential of the proposed method. The proposed method can be used for not only communities of bacteria, but also for communities of people, such as sports teams and classes in school. In the present paper, the terms "group" and "member" are used to explain the hierarchical structures of relational data. Therefore, the proposed method can be widely applied to relational data with such a hierarchical structure.

4. Conclusion

We have addressed the problem of extrapolating group proximities from member relations (group proximity problem) and proposed a method based on embedding and distribution mapping. In the present paper, a metric MDS and the SOM^2 were used for embedding and distribution mapping, respectively. The proposed method was applied to the NIPS document and bacterial flora datasets. The results of these experiments confirmed that the proposed method is useful for solving the group proximity problem. In the future, we intend to investigate the effects of the non-Euclideanity and preservation accuracy on group proximities and to compare the proposed method with other approaches for solving the group proximity problem.

References

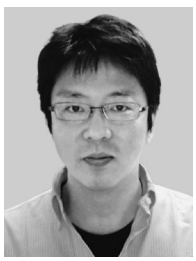
[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," J. Educ. Psychol., vol.24, no.6, pp.417–441, 1933.
[2] T. Kohonen, Self-Organizing Maps, 3rd ed., Springer-Verlag, Berlin,

2001.
[3] E. Pekalska and R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition: Foundations and Applications, World Scientific, Singapole, 2005.
[4] T.F. Cox and M.A.A. Cox, Multidimensional Scaling, Chapman & Hall, London, 1994.
[5] I. Borg and P. Groenen, Modern Multidimensional Scaling: Theory and Applications, 2nd ed., Springer, New York, 2005.
[6] T. Furukawa, "SOM of SOMs," Neural Netw., vol.22, no.4, pp.463–478, 2009.
[7] J.B. Tenenbaum, V. de Silva, and J.C. Langford, " A global geometric framework for nonlinear dimensionality reduction," Science, vol.290, no.5500, pp.2319–2323, 2000.
[8] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol.290, no.5500, pp.2323–2326, 2000.
[9] G.E. Hinton and S.T. Roweis, "Stochastic neighbor embedding," in Advances in Neural Information Processing Systems 15, pp.833–840, MIT Press, Cambridge, MA, 2002.
[10] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," J. Mach. Learn. Res., vol.8, pp.2265–2295, 2007.
[11] A. Ultsch and H.P. Siemon, "Kohonen's self organizing feature maps for exploratory data analysis," Proc. Int. Neural Network Conf., pp.305–308, Dordrecht, Netherlands, 1990.
[12] N. Morotomi, K. Fukuda, M. Nakano, S. Ichihara, S. Oono, T. Yamazaki, N. Kobayashi, T. Suzuki, Y. Tanaka, and H. Taniguchi, "Evaluation of intestinal microbiotas of healthy Japanese adults and effect of antibiotics using the 16S ribosomal RNA gene based clone library method," Biol. Pharm. Bull., vol.34, no.7, pp.1011–1020, 2011.
[13] T. Kawanami, K. Fukuda, K. Yatera, T. Kido, C. Yoshii, H. Taniguchi, and M. Kido, "Severe pneumonia with Leptotrichia sp. detected predominantly in bronchoalveolar lavage fluid by use of 16S rRNA gene sequencing analysis," J. Clin. Microbiol., vol.47, no.2, pp.496–498, 2009.
[14] T. Smith and M. Waterman, "Identification of common molecular subsequences," J. Mol. Biol., vol.147, no.1, pp.195–197, 1981.
[15] W.R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms," Genomics, vol.11, no.3, pp.635–650, 1991.

Hideaki Misawa received the B.E. degree from the National Institution for Academic Degrees and University Evaluation, Japan, in 2005 and M.E. degree from Kyushu Institute of Technology, Japan, in 2007. He is currently a Ph.D. student at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan. His research interests include soft computing technology and its application to the biomedical field. He is a student member of IEEE.

Keiichi Horio received the B.E., M.E., and Ph.D. degrees from Kyushu Institute of Technology, Japan, in 1996, 1998, and 2001, respectively. He is currently an associate professor at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan. His research interests include soft computing technology and its applications. He is a member of the IEEE, SOFT, and BMFSA.

Nobuo Morotomi received the M.D. degree in 2002. He is currently an adjunct research assistant of the Department of Microbiology at the University of Occupational and Environmental Health, Japan. His research interests include evaluation of intestinal microbiota.



Kazumasa Fukuda received the Ph.D. degree from Nagasaki University, Japan, in 1998. He is currently an assistant professor of the Department of Microbiology, School of Medicine, University of Occupational and Environmental Health, Japan. His research interests include medical bacteriology and environmental microbiology.

Hatsumi Taniguchi is a professor of the Department of Microbiology, School of Medicine, University of Occupational and Environmental Health, Japan. Her research interests include molecular analysis of recurrent mechanism of Mycobacterium spp. and culture-independent molecular analysis of bacterial flora.