2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

A Study on Extraction of Minority Groups in Questionnaire Data based on Spectral Clustering

Kazuto Inagaki
Nagoya University
Email: inagaki@cmplx.cse.nagoya-u.ac.jp

Tomohiro Yoshikawa
Nagoya University
Email: yoshikawa@cse.nagoya-u.ac.jp

Takeshi Furuhashi
Nagoya University
Email: furuhashi@cse.nagoya-u.ac.jp

***Abstract*—In the field of marketing, a questionnaire is one of the most important approaches in order to research the market or to design a marketing strategy. On the other hand, people have a variety of individuality recently, then respondents have various impressions on evaluation objects. In the analysis of collected questionnaire data, it is important not only to analyze overall trends but also to discover minority groups which have strong impressions but are different from general groups. It is, however, difficult to extract minority groups by conventional cluster analysis applied to questionnaire data, because they generally aim at extracting majority groups or making a rough clustering. In this paper, we propose the extraction method of minority groups in questionnaire data using the spectral clustering method which considers local similarity and extracts the clusters having less connection to general groups.**

I. Introduction

In marketing, it is very important that companies grasp customers' impression on their products and services through market research. For example, when a company develops a new product, it designs a marketing strategy after understanding the target customers' demands and impressions to ready-made products[1], [2].

One of the methods for the market research is a questionnaire by the rating scale method[3] or the semantic differential method[4], and it can obtain the questionnaire data in which people's impressions on evaluation objects are quantified by answering a set of questions for each object with multiple grading scales. The obtained questionnaire data is generally analyzed using the multivariate analysis methods such as cluster analysis[5], principal component analysis (PCA)[6], multidimensional scaling method (MDS)[7], etc. These approaches, however, often aim at analyzing the overall trends and characteristics, and they regard the answers which differ from the overall trends greatly as noises which may give a negative effect to the analysis result. That makes it difficult to extract the groups which are a small number but have the interesting characteristics on analysis called "minority groups."

In this paper, we try to extract these minority groups in respondents using spectral clustering[8]. Spectral clustering is the clustering method by a graph partitioning, and it clusters data whose similarities in a subgraph are high and those with other subgraphs are low. Therefore, it is thought that this method is suitable for the extraction of the minority groups which are small but have strong impressions different from others.

In this paper, we propose a method which defines the similarity between respondents based on a gaussian function and extracts minority groups one by one by the iteration of division into two groups. Furthermore, we also propose a method to determine the parameter automatically in the similarity function between respondents based on Bayesian Information Criterion. First, it is shown that the assumed minority groups can be extracted appropriately by the proposed method in the preliminary experiment with a virtual questionnaire data comparing with the conventional method. Next, the proposed method is applied to an actual questionnaire data, and it is shown that some groups of a small number of respondents with different characteristics from the average are extracted.

II. Spectral Clustering

Spectral clustering is the method of solving a clustering as a problem of graph partitioning by regarding a data as a node and the similarity between data as the weight of edge between nodes. The whole graph expressed in this way is divided into some subgraphs by cutting some edges. An evaluation function which makes the edge in a subgraph dense and that between subgraphs sparse is defined. Although some evaluation functions have been proposed, $Ncut$, which is the representative function, is employed in this paper. Giving a partition on V, the nodes of a graph, and dividing it into two subgraphs A and B, $cut(A, B)$, the similarity between subgraph A and B, is defined as follows.

$$cut(A, B) = \sum_{i \in A, j \in B} w(i, j) \tag{1}$$

where $w(i, j)$ is the weight of the edge between node i and j. Then the evaluation function $Ncut$ is expressed in the following equation.

$$Ncut(A, B) = \frac{cut(A, B)}{cut(A, V)} + \frac{cut(A, B)}{cut(B, V)} \tag{2}$$

It is equivalent to make similarities in a subgraph large and those between subgraphs small to minimize this function. It is known that this minimization problem will result in a generalization eigenvalue problem. When W is a similarity matrix and D is a matrix which has a degree of W in the diagonal component, the eigenvector of $D^{-1}(D - W)$ gives the division of a graph. Since the smallest eigenvalue is set to 0, the second smallest eigenvector is used. The nodes which have an element value more than a certain value are assigned to cluster A, and those have an element value less than a certain value are assigned to cluster B. The threshold corresponding to

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

the element value is mainly set to 0, median, or the value which minimizes $Ncut$. In this paper, we calculate $Ncut$ cutting in each point and determine the value which minimizes $Ncut$.

III. Proposed Method

This section describes the extraction method of minority groups using spectral clustering described in the previous section. Here, we define "minority group" as a small group of respondents who answer differently from others but similarly one aother.

A. Similarity Definition

Given the vectors $\boldsymbol{x_a}$,$\boldsymbol{x_b}$, which have the scores for questions by respondent a and b, the similarity between them is defined by the following equation.

$$w(a,b) = \exp(-\frac{||\boldsymbol{x_a} - \boldsymbol{x_b}||^2}{\sigma^2}) \quad (3)$$

Equation (3) is called gaussian function and σ^2 is a parameter representing a variance value. This function emphasizes the similarity between $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ when $||\boldsymbol{x_a} - \boldsymbol{x_b}||$ is small and makes it approximately 0 when $||\boldsymbol{x_a} - \boldsymbol{x_b}||$ is large. The smaller σ is, the more extreme this emphasis becomes.

As described in the section II, spectral clustering divides a graph so that the similarities in a subgraph becomes large and those between subgraphs becomes small. Therefore, by defining the similarity between respondents as eq. (3), the similarity in a group and dissimilarity with others are emphasized, and it is expected that minority groups required in this study can be extracted appropriately.

B. Determination of Parameter σ^2

σ^2 in eq. (3) should be determined before the clustering. However, it is difficult to determine the proper value of σ^2 which effects the result of clustering greatly. Usually, it is necessary to grasp the characteristics of data from multiple different perspectives by iteration of trial and error in the analysis of questionnaire data. Therefore, it can be also one of the effective approaches to analyze the acquired group respectively by varying the value of σ^2. In this paper, however, we propose a method to determine σ^2 automatically based on the assumption that a minority group follows a multivariate normal distribution which is dense locally. In [9], the X-means method is proposed as the decision method of number of clusters using Bayesian Information Criterion (BIC)[10] in the K-means method which is one of the representative clustering methods. BIC is expressed by the following equation.

$$BIC = -2 \log L + k \log n \quad (4)$$

where L is a likelihood, n is a sample size, and k is the number of population parameters. In the proposed method, we change the value of σ^2 within a range, calculate BIC to the multivariate normal distribution of the minority group extracted with each σ^2 value, and determine the value of σ^2 which minimizes BIC.

C. Extraction of Minority Groups by Repetition of Two Division

Spectral clustering is extended to the division into more than two clusters[11]. However, this method needs to determine the number of clusters in advance, so the application to the analysis of questionnaire data is difficult because the number of the existing minority groups is unknown. Therefore, the proposed method extracts minority groups one by one by the iteration of two division described in the section II to the cluster of the maximum number. It is considered to be possible that conventional methods also extract minority groups when the number of clusters is large enough. In this case, however, we need to search the characteristic minority groups from a lot of acquired clusters. Therefore, the extraction of minority groups and the analysis of them one by one can be practical.

D. Algorithm

The algorithm of the proposed method is described below.

Algorithm 1 Algorithm of proposed method

```
set of all respondents $X$
set of targeted respondents $X'$
search region of $\sigma^2$ $region_{\sigma^2}$

Require: $region_{\sigma^2}$
Ensure: set of minority groups $M$
  $X' \Leftarrow X$
  repeat
    for all $\sigma^2$ in $region_{\sigma^2}$ do
      divide $X'$ into $A$ and $B$ by Spectral Clustering with $\sigma^2$, subject to $\#(A) \leq \#(B)$
      calculate $BIC$ of $A$
    end for
    set $A$ whose $BIC$ is minimized as minority group $A_M$
    $X' \Leftarrow X' \setminus A_M$
    add $A_M$ to $M$
  until $M$ becomes enough for analyzer
```

IV. Related Work

Basically, there are few reports of research aiming at an extraction of minority groups in the analysis of questionnaire data. Some methods can be applied to extract minority groups. However, most of them mainly aim at extracting outliers and abnormal data rather than minority groups. Therefore, when we apply these method to actual questionnaire data, abnormal data will be extracted one by one in many cases[12], [13], [14], [15].

Ando *et. al.* proposed a clustering method for mixed data in which majority groups distributed globally and minority groups distributed locally using the information theoretical clustering in order to detect minority groups[16]. Gonzalez *et. al.* proposed an extraction method of dense data distributed locally by the iteration of weak clusterings[17] with low calculation cost[18]. However, these methods tend to extract the densest groups in a distribution. Questionnaire data generally have a lot of respondents who answer with scores around the median value to all questions. Therefore, these respondents of a majority group without any interesting character will be extracted as a minority group by these methods. As described above, this study aims at extracting small groups with high similarity with inside and low similarity with outside. Therefore, it is difficult to extract minority groups by the above methods. Furthermore, Ando's method[16] requires the assumption of the distributions for both majority and minority groups beforehand.

Fukami *et. al.* proposed a method to extract minority groups by the visualization based on the data configuration error by MDS[19]. The proposed method differs in terms of the presupposition of this method that finds minority groups by trial and error with the iteration of grouping respondents manually.

V. Experiments

In this section, we applied the proposed method described in the section III to a virtual and an actual questionnaire data, respectively, and evaluated the performance of the proposed method comparing with the conventional method.

A. Application to Virtual Questionnaire Data

1) Virtual Questionnaire Data: We generated a virtual questionnaire data on a five-step scale method with 1 evaluation object, 10 questions, and 650 respondents. The respondents were classified into seven groups shown in TABLE I. In TABLE I, Group 5, 6, and 7 were the assumed minority groups in this experiment.

TABLE I. Characteristics of each group in virtual questionnaire data

Group	Number of Respondents	Characteristics
Group1	100	Randomly marking 4 or 5 to the question 1-3, and 1 or 2 to the question 4-10.
Group2	300	Randomly marking 3-5 to the question 1-3, and 1-3 to the question 4-10.
Group3	100	Randomly marking 2-4 to all the question.
Group4	100	Randomly marking 1-5 to all the question.
Group5	10	Randomly marking 1 or 2 to the question 1-3, and 4 or 5 to the question 4-10.
Group6	20	Randomly marking 4 or 5 to all the question.
Group7	20	Randomly marking 1 or 2 to all the question.

Figure 1 shows the distribution of the respondents' scores by MDS based on Euclidean distance between 10-dimensional vectors which consist of the scores of 10 questions as elements. Figure 2 shows the average scores of each group and all respondents. In Fig. 1, the distortion due to the dimension compression by MDS visually divided Group7 into two parts.

2) Experimental Setup: We extracted minority groups three times to the virtual questionnaire data described above by the proposed method as a preliminary experiment. In each extraction, we determined the value of σ^2 in the range from 1 to 10 with 0.2 unit by the method described in the section III-B. Moreover, we compared the result of the proposed method with that of dendrogram[5] which was one of the representative cluster analysis methods.

3) Results and Discussions: Figure 3 shows the visualization result of three clusters extracted by the proposed method, and Fig. 4 shows the average scores of each cluster. Figure 5 shows the value of BIC calculated by eq. (4) to each σ^2 of each extraction.

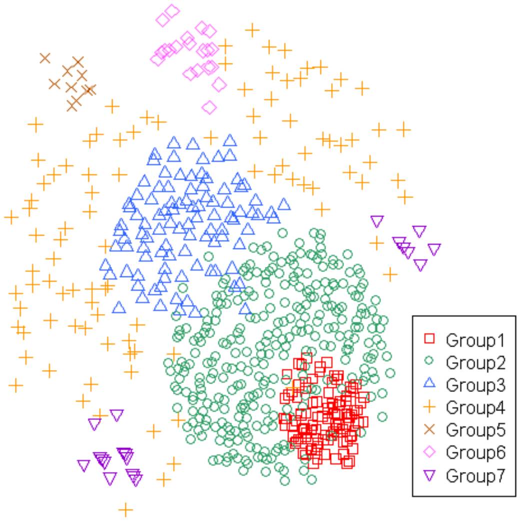
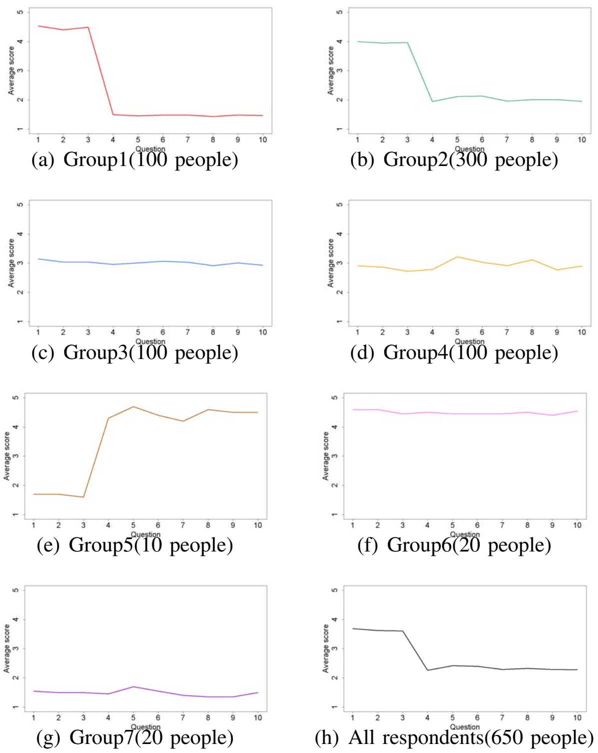


Fig. 1. Visualization of evaluation scores



(a) Group1(100 people)
(b) Group2(300 people)
(c) Group3(100 people)
(d) Group4(100 people)
(e) Group5(10 people)
(f) Group6(20 people)
(g) Group7(20 people)
(h) All respondents(650 people)

Fig. 2. Number of respondents and average scores of each group (virtual data)

The proposed method appropriately extracted Group 5, 6, and 7 which were the assumed minority groups in order of the Group 6, 5, and 7, respectively. As shown in Fig. 4, one more respondent was clustered together comparing to the set-up number in Group 5 and 7. That was because there were the respondents whose scores were similar to those of Group 5 and 7, respectively, in Group 4, the scores were marked randomly to all questions. Conversely, the respondents of Group 5, 6, and

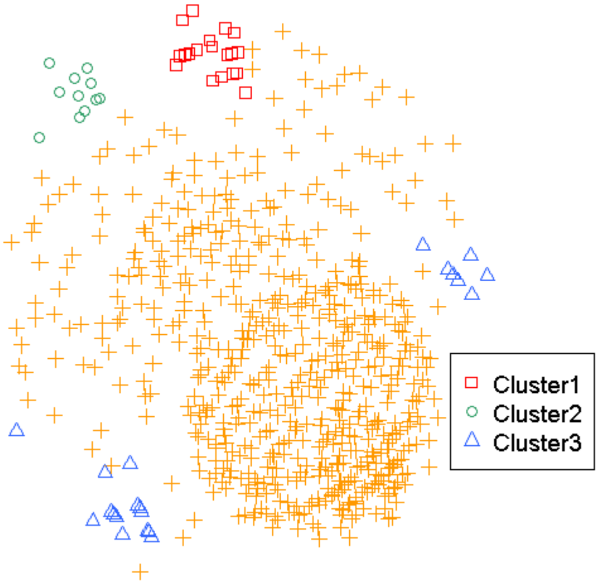
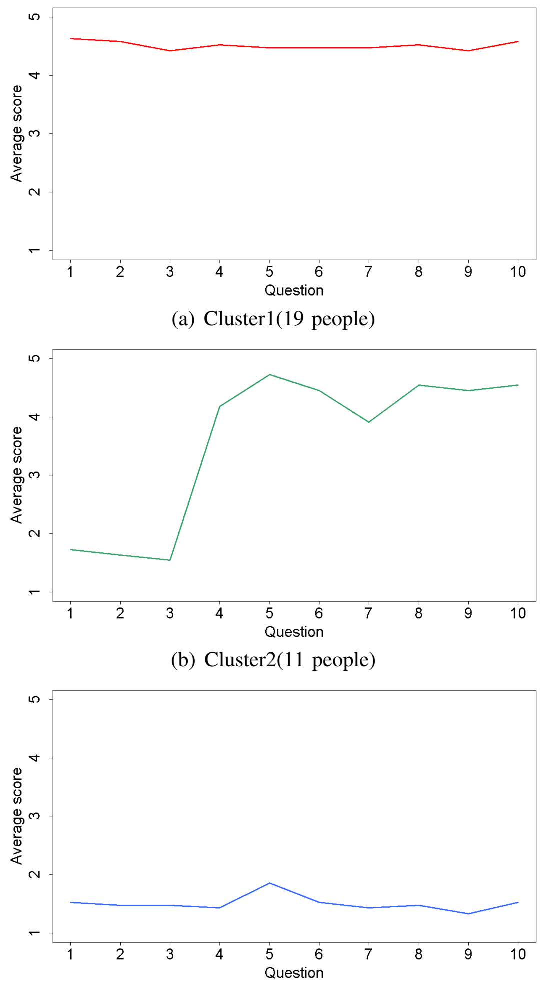


Fig. 3. Clustering result by proposed method (virtual data)

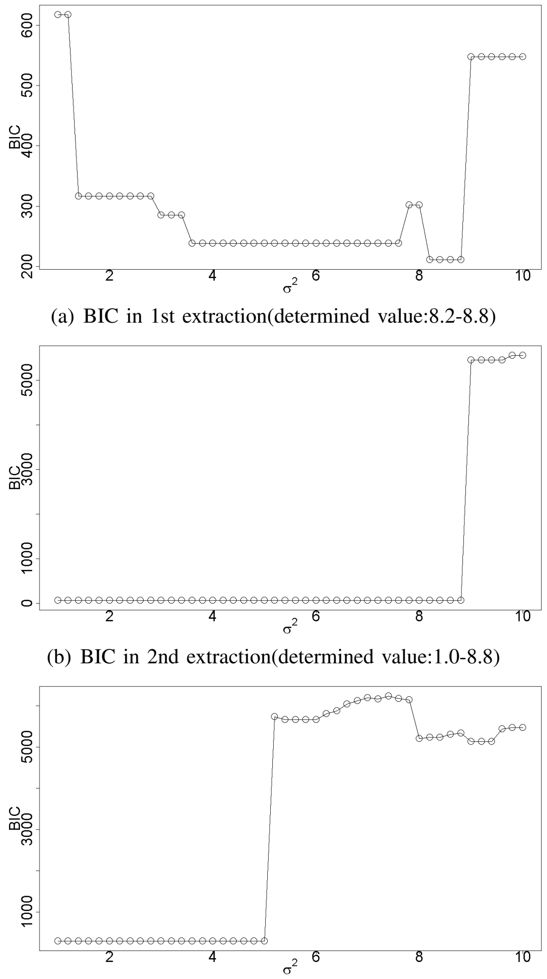


(a) Cluster1(19 people)

(b) Cluster2(11 people)

(c) Cluster3(21 people)

Fig. 4. Average scores of extracted clusters (virtual data)



(a) BIC in 1st extraction(determined value:8.2-8.8)

(b) BIC in 2nd extraction(determined value:1.0-8.8)

(c) BIC in 3rd extraction(determined value:1.0-5.0)

Fig. 5. Results of σ^2

7 who were assumed as minorities were altogether covered by Cluster 2, 1, and 3 except for one respondent in Group 6. This result shows that the proposed method could extract minority groups appropriately. Moreover, Fig. 5 showed that the optimal values of σ^2 were determined in wide ranges.

Figure 6 shows the result by dendrogram. We employed Ward's method[20] for the distance between clusters because this method is said to have good performance in classifications. Figure 6 shows that Group 5, 6, and 7 were clustered respectively and relatively separated from other groups. However, when the respondents are divided hierarchically, Groups 7, 6, and 5 will be extracted in the 3rd, 4th, and 6th division. It is difficult to extract only these groups as minorities alone. Ward's method has a tendency to make the number of each group equal because it clusters according to the criterion of maximizing variance between groups to that in a group. On the other hand, the single connecting method[21] is introduced as a method of permitting the deviation of the number of each cluster. However, it tends to combine the data one by one with clusters, which is called a chain. Therefore, this method is not suitable for extracting minority groups.

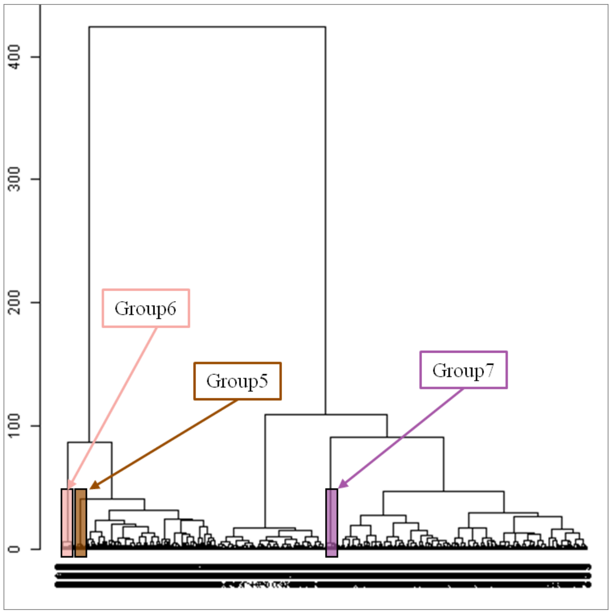


Fig. 6. Result of dendrogram (virtual data)

B. Application to Actual Questionnaire Data

1) Actual Questionnaire Data: A Web questionnaire survey was carried out involving 1,014 respondents for next-generation services as the evaluation objects. In the questionnaire, the rating scale method was employed and the respondents were asked to choose one of five grades 1, 2, 3, 4, 5 in the responses of 10 questions. In this survey, grade 5 means "strongly agree" while grade 1 means "strongly disagree." Table II shows the 6 next-generation services used as the evaluation objects and Table III shows the 10 questions for every object. Note that the evaluation objects were shown to respondents by more concrete description of them in TABLE II in the actual questionnaire.

TABLE II. EVALUATION OBJECTS

Object	Content
Object1	Unclear explanation about after-sale service
Object2	Unclear explanation about ubiquitous
Object3	Unclear explanation about recycle
Object4	Detailed explanation about after-sale service
Object5	Detailed explanation about ubiquitous
Object6	Detailed explanation about recycle

TABLE III. QUESTIONS

Question	Content
Question1	I'm interested in what it is.
Question2	I want to recommend it to people around.
Question3	It has a high social demand, likely to spread.
Question4	The image of the companies providing it is likely to improve.
Question5	The burden of the companies providing it is too large.
Question6	The aim is wrong.
Question7	It would be appreciated only by certain people.
Question8	It has the essence of social issues.
Question9	Although it is socially important, public authorities should assist because the burden of the companies providing it is large.
Question10	It is the futuristic service.

2) Experimental Setup: We extracted minority groups by the proposed method like the section V-A2. A 60-dimensional vector to 10 questions for 6 objects was used for the score vector of each respondent. We determined the value of σ^2 in the range from 1 to 10 with 0.5 unit. Clustering of respondents was also done by dendrogram.

3) Results and Discussions: Figure 7 shows the visualization of the respondents' scores by MDS with Cluster 1-5 extracted by the proposed method. Figure 8 shows the number of respondents and the average scores of each cluster.

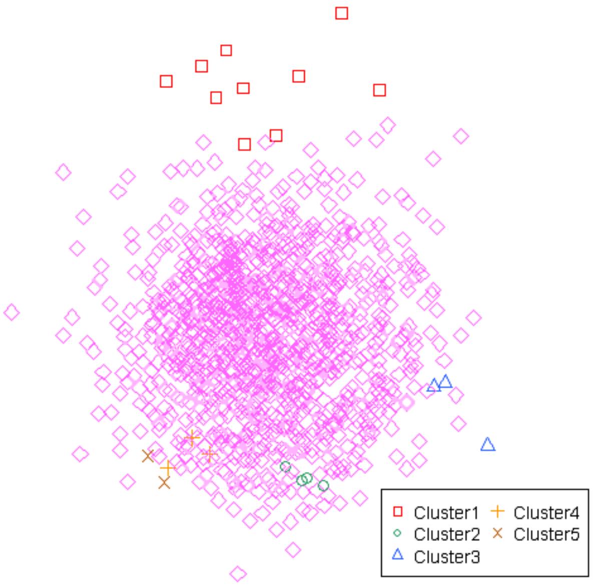
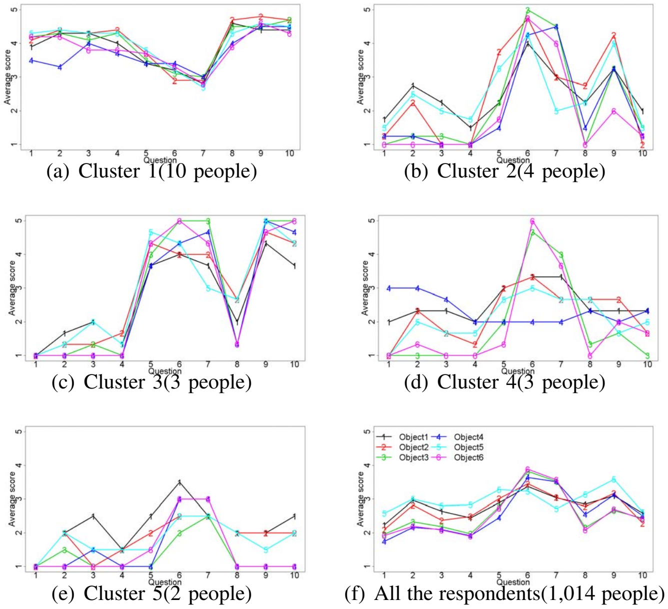


Fig. 7. Clustering result by proposed method (actual data)



(a) Cluster 1(10 people)

(b) Cluster 2(4 people)

(c) Cluster 3(3 people)

(d) Cluster 4(3 people)

(e) Cluster 5(2 people)

(f) All the respondents(1,014 people)

Fig. 8. Number of respondents and average scores of each cluster (actual data)

Cluster 1 in Fig. 8(a) was the respondents whose tendency of scores was opposite to that of all respondents in Fig. 8(f). The average scores of Cluster 2 in Fig. 8(b) were 1 or 5 in every question, and it was found that they answered using extreme scores relatively. Clusters 3 and 4 were similar to Cluster

2, while the difference between them was the difference of scores to Question 9 and 10. Cluster 5 was the respondents answering by low scores to almost all questions. Thus, it is considered that the proposed method could also extract characteristic minority groups in the actual questionnaire data. Moreover, the result of dendrogram in Fig. 9 shows that the extraction of characteristic groups considered to be minority groups is difficult.

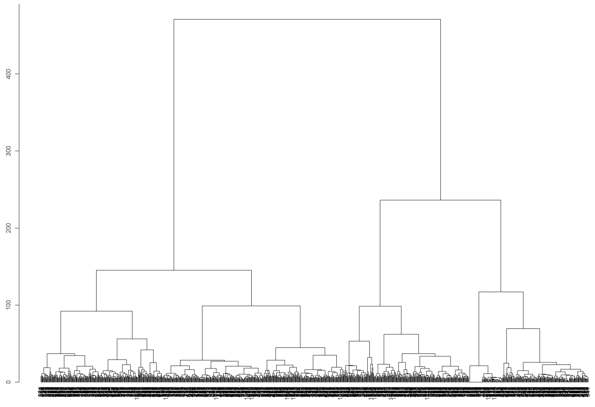


Fig. 9. Result of dendrogram (actual data)

VI. Conclusion

In this paper, we proposed the extraction method of minority groups in questionnaire data based on the spectral clustering. First, it was shown that the proposed method could extract the assumed minority groups appropriately in the virtual questionnaire data as a preliminary experiment. Next, we applied the proposed method to the actual questionnaire data, and it was shown that some groups of a small number of respondents with different characteristics from the trends of other respondents were extracted. As future work, we will investigate the validity of extracted minority groups and analyze the relationship of the similarity function between respondents and the obtained result.

References

[1] S. Kuroda, K. Yamamoto, T. Yoshikawa, and T. Furuhashi, "A proposal for analysis of sd evaluation data by using clustering method focused on data distribution," in *Frontiers of Computational Science*, pp. 317–320, Springer, 2007.

[2] M. Futatsuka, J. Yonesaki, M. Ikeda, *et al.*, "Relationship between questionnaire survey results of vibration complaints of wheelchair users and vibration transmissibility of manual wheelchair," *Environmental health and preventive medicine*, vol. 8, no. 3, pp. 82–89, 2003.

[3] C. E. Osgood, *The measurement of meaning*, vol. 47. University of Illinois Press, 1957.

[4] S. H. Hsu, M. C. Chuang, and C. C. Chang, "A semantic differential study of designers' and users' product form perception," *International Journal of Industrial Ergonomics*, vol. 25, no. 4, pp. 375–391, 2000.

[5] M. R. Anderberg, "Cluster analysis for applications," tech. rep., DTIC Document, 1973.

[6] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[7] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[8] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[9] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters.," in *ICML*, pp. 727–734, 2000.

[10] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[12] B. Wang, G. Xiao, H. Yu, and X. Yang, "Distance-based outlier detection on uncertain data," in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*, vol. 1, pp. 293–298, IEEE, 2009.

[13] E. M. Knox and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the International Conference on Very Large Data Bases*, Citeseer, 1998.

[14] E. M. Knorr, *Outliers and data mining: finding exceptions in data*. PhD thesis, The University of British Columbia, 2002.

[15] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[16] S. Ando and E. Suzuki, "Detecting clusters of outliers with information theoretic clustering," *Transactions of the Japanese Society for Artificial Intelligence(in Japanese)*, vol. 23, pp. 344–354, 2008.

[17] A. Topchy, A. K. Jain, and W. Punch, "Combining multiple weak clusterings," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 331–338, IEEE, 2003.

[18] E. Gonzàlez and J. Turmo, "Unsupervised ensemble minority clustering," *Machine Learning*, pp. 1–52, 2012.

[19] T. Fukami, Y. Watanabe, T. Yoshikawa, T. Furuhashi, I. Hara, and H. Yoneda, "Discovering minority groups by interactive clustering in visible space," in *International Conference on Kansei Engineering and Emotion Reserch*, vol. 2009, 2009.

[20] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[21] J. C. Gower and G. Ross, "Minimum spanning trees and single linkage cluster analysis," *Applied statistics*, pp. 54–64, 1969.