VISUALIZATION OF MULTIVARIATE DATA USING EXPANDED CONSTELLATION AND EXPANDED KANJI GRAPHS AND THEIR APPLICATION TO CLUSTERING

Mika FUJIWARA1*, Shoji KAJINISHI2 and Koji KURIHARA3*

1 Graduate School of Environmental and Life Science, Okayama University
(3-1-1,Tsushima-Naka, Kita-ku, Okayama 700-8530, Japan)
* Corresponding author, E-mail: mika.f@cjc.ac.jp
2 Graduate School of Environmental and Life Science, Okayama University
(3-1-1,Tsushima-Naka, Kita-ku, Okayama 700-8530, Japan)
E-mail: skajinishi@s.okayama-u.ac.jp
3 Professor, Graduate School of Environmental and Life Science, Okayama University
(3-1-1,Tsushima-Naka, Kita-ku, Okayama 700-8530, Japan)
* Corresponding author, E-mail: kurihara@okayama-u.ac.jp

In this study, expanded constellation and expanded kanji graphs are proposed and the effectiveness of these graphs in the multivariate analysis is examined. To draw the expanded constellation graph, the variables are first placed on the circumference of the semicircle using factor loadings. Then, a line is drawn inside the semicircle by connecting the vectors of each variable. The constellation graph can simultaneously grasp both the tendencies of the objects and p-dimensional variables. In the expanded kanji graph, the height and width of the kanji are determined using the mean values of the variables for each cluster. The kanjis are placed around the radar chart drawn using the standard deviation. The kanji graph can intuitively grasp the characteristics of each group in the cluster analysis.

Herein, the effectiveness of the proposed methods is verified by evaluating the flavor of whiskey. The proposed methods show that whiskey flavors can be classified into five clusters: (1) “full-body and winey type,” (2) “sweet and balanced type,” (3) “smoky and balanced type,” (4) “full-body and smoky type,” and (5) “light-body and sweet floral type.”

Key Words : graph representation, factor loading, cluster analysis, expanded constellation graph, expanded kanji graph

1. INTRODUCTION

In statistical analysis, graph representations are used to capture the characteristics of the data to facilitate the interpretation of the graph’s results. The graph representation is a visualization method for intuitively grasping the summary of information and detailed features of the data, and it is effective for achieving the exploratory interpretation of the results. Typical graph representations for visually drawing the information and features of data include the circle graph, bar graph, scatter diagram, and radar chart. Besides these graphs, various graph methods such as face graph1, Andrews graph2, tree graph3, constellation graph4, and kanji graph5 have been proposed. These graph representations are easy to derive and calculate. Moreover, they can easily and intuitively interpret data features.

Herein, an expanded constellation graph was proposed, which is a new graph that uses the concept of constellation graphs. To complement the expanded constellation graph, an expanded kanji graph was also proposed, which adopts the concept of kanji graphs. The usefulness of these methods was verified by employing them to evaluate the flavor data of whiskey.
2. PROPOSED METHOD

(1) Expansion of the constellation graph

For each object, a constellation graph\(^\text{11}\) represents each variable as a vector and connects these vectors and draws a star at the final point. The concatenated vectors are called “path.” The constellations of all the data are always drawn inside the semicircle. The constellation graph can grasp the mean values of the variables and the variation between the variables. Moreover, it can intuitively grasp the characteristics of the variables of the object from the shape of the path.

For this graph, let the data be represented as \(x_{ij} (i = 1, 2, \cdots, n, j = 1, 2, \cdots, p)\) with \(p\) variables. Each variable is transformed and expressed as a vector as follows.

\[
\theta_{ij} = f_j(x_{ij}) \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, p, \quad 0 \leq \theta_{ij} \leq \pi
\]

The following equation can be considered as an example of conversion.

\[
\theta_{ij} = \frac{x_{ij} - L_j}{U_j} \pi,
\]

where \(U_j = \max_{1 \leq i \leq n} x_{ij}\) and \(L_j = \min_{1 \leq i \leq n} x_{ij}\).

Using this \(\theta_{ij}\), the coordinates \((\alpha, \beta)\) of the final point of the constellation graph are determined as follows.

\[
(\alpha, \beta) = \left( \sum_{j=1}^{p} w_j \cos \theta_{ij}, \sum_{j=1}^{p} w_j \sin \theta_{ij} \right)
\]

However, the weight \(w_j\) that determines the length of the vector should satisfy the following conditions to ensure that the final point falls within the semicircle.

\[
\sum_{j=1}^{p} w_j = 1; \quad (w_j \geq 0, \quad j = 1, 2, \cdots, p)
\]

Here, the larger the variation between the variables, the closer is the final point of the star to the circumference of the semicircle. Where, \(1/p\) is generally used for \(w_j\). Moreover, as shown in reference 4, the smaller the mean value, the more the final point will be placed to the right.

An expanded constellation graph is proposed herein. This graph reflects the value of the factor loading obtained using the factor analysis and places the variables along the circumference of the semicircle. Furthermore, in the graph, the factor loading is assigned to the direction of the vector and the data value is assigned to the length of the vector. This graph is expected to facilitate the interpretation of the results by visualizing both the tendency of factors and the values of variables based on factor analysis.

In the expanded constellation graph, factor analysis is performed using the data \(x_{ij} (i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, p)\), with \(p\) variables. The factor loading \(a\) of the first factor obtained using this analysis is sorted in descending order. At this moment, the maximum and minimum values of \(a\) are \(a_u\) and \(a_l\), respectively. Next, each variable is represented as a vector using the following transformation.

\[
\varphi_j = \frac{a_u - a_j}{a_u - a_l} \pi; \quad a_u = \max a_i, \quad a_l = \min a_j
\]

When the constellation graph of the \(i\)-th data is drawn using this \(\varphi_j\), the coordinates \((\alpha, \beta)\) of the final point are determined as follows.

\[
(\alpha, \beta) = \frac{1}{p} \left( \sum_{j=1}^{p} \frac{x_{ij}}{U_j} \cos \varphi_j, \quad \sum_{j=1}^{p} \frac{x_{ij}}{U_j} \sin \varphi_j \right); \quad (x_{ij} > 0), \quad \text{where } x_{ij} \leq 0 \text{ transformed such that } x_{ij} > 0.
\]

By connecting the vectors in a similar manner for each object and drawing a star at the final point, the path and star of each object are drawn in the semicircle.

After this procedure, a graph is drawn for each factor. As mentioned above, in existing constellation graphs, the length of the vector is common to all variables and the direction \(\theta_{ij}\) depends on the value of the data. However, in the expanded constellation graph, the direction is common to all variables, and the length of the vector depends on the raw score. Therefore, in the expanded constellation graph, the position of the final point is affected by the order of the data. Owing to this difference, in the existing constellation graph, the result derived from the data is fixed. However, the expanded constellation graph can be expressed in different dimensions such as the first and second factors and interpreted from a wide range of perspectives.

(2) Expansion of the kanji graph

The kanji graph visualizes each variable by assigning the raw score of the variable to the size of a single kanji. In the kanji graph, one variable is represented by single kanji. This graph can express the properties of the variables using kanjis, and the characteristics of the variables in each object can be intuitively grasped by comparing kanji sizes.

The procedure for drawing a kanji graph is shown below.

1) The data \(x_{ij} (i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, p)\) consists of \(n\) observations and \(p\) variables. The following transformations is performed using \(x_{ij} (i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, p)\):

\[
x_{ij}' = \frac{x_{ij} - \bar{x}_j}{S_j} + b,
\]

where \(b\) is a positive constant that satisfies \(x_{ij}' \geq 0\) at all times. \(\bar{x}_j\) is the mean of variable \(X_j\), and \(S_j\) is object standard deviation of variable \(X_j\).

2) Single kanji that represents the characteristics of the variable is selected arbitrarily. Because a variable
can be represented using a kanji, \( p \) types of kanji must be prepared for \( p \) variables.

3) The height and width of the kanji are determined according to the values of \( x_{ij} \).

In object \( i \), the height and width of the first kanji are transformed as

\[
\ell \left( \frac{x_{i1}}{b} \right),
\]

where \( \ell \) is a size of the kanji.

Similarly, the \( p \)-th kanji is drawn using the values of \( x_{ip} \).

4) Kanjis are drawn and arranged from left to right.

5) All objects are drawn using this procedure.

As mentioned above, the existing kanji graph represents a kanji string for each object. However, when drawing a kanji graph for a cluster consisting of multiple objects, it is necessary to use the mean or median value of each variable in the cluster. Additionally, the interpretation of results differs depending on the degree of dispersion indicators, such as standard or quartile deviation, even if the mean value is the same.

In the proposed expanded kanji graph, a radar chart is created using the standard deviation (\( S_j \)) of each variable and the kanjis are placed around the circumference of the radar chart. Generally, in this graph, the position of the final point depends on the order of the variables. The order of variables can be arbitrarily decided by the analyst, or using the results of factor analysis etc. This graph clarifies the characteristics of each group by visualizing the mean value and standard deviation of each group. Furthermore, by visualizing the variables using the kanji, the properties of the variables can be intuitively interpreted.

The procedure of drawing an expanded kanji graph is described below.

1) The order of the variables placed on the radar chart is arbitrarily determined.

2) A radar chart is created using \( S_j \).

3) \( p \) kanjis that match the properties of each variable are selected.

4) The height and width of the kanji are determined according to the mean value of each variable using the same procedure as that used for the existing kanji graph. When all variables have the same scale, the kanji size is determined by multiplying the value of the variable by an arbitrary magnification \( \ell \).

5) The kanjis with the height and width adjusted are placed on the radar chart.

3. APPLICATION EXAMPLE OF THE METHOD

(1) Data used

In this section, to confirm the usefulness of the proposed methods, whiskey data provided by the University of Strathclyde in Scotland were evaluated. The evaluation of 12 flavors (medicinal, smoky, body, spicy, winey, nutty, malty, honey, fruity, sweetness, floral, and tobacco) of 86 brands of whiskey was performed. The 12 flavors were rated on a 5-point scale (0–4 points).

First, the data set was subjected to factor analysis and the whiskey brands data were classified using the k-means method. Next, to examine the effectiveness of each method, the expanded constellation and expanded kanji graphs were employed to evaluate the five clusters obtained using the k-means method.

For statistical analysis, R version 3.5.1 and IBM SPSS Statistics for Windows version 26 were used.

(2) Interpretation using the expanded constellation graph

First, the 12 flavors were subjected to exploratory factor analysis (maximum likelihood method/promax rotation). The analysis was repeated under the commonality criterion of .20 or more, and nine flavors and two factors were extracted based on the Gatman Kaiser’s criterion (eigenvalue of 1.0 or more), as shown in Table 1. The contribution rates of factors 1 and 2 were 29.4% and 14.1%, respectively.

| Item       | Factor1 | Factor2 | Commonality |
|------------|---------|---------|-------------|
| Medicinal  | .911    | -.128   | .846        |
| Smoky      | .739    | .029    | .547        |
| Tobacco    | .534    | -.042   | .287        |
| Body       | .360    | .638    | .537        |
| Spicy      | -.013   | .065    | .004        |
| Winey      | -.259   | .733    | .604        |
| Sweetness  | -.406   | .171    | .194        |
| Honey      | -.514   | .368    | .400        |
| Floral     | -.596   | -.346   | .475        |

Table 1 Results of factor analysis.

The first factor was named “smoky flavor vs. delicious flavor” because the factor load on the variables of medicine, smoky, and tobacco is high in the positive direction and the factor load on the variables of sweetness, honey, and floral is high in the negative direction. The second factor was named “depth and winey flavor” because the factor load on
the variables of body and winey was high. Next, an expanded constellation graph was drawn based on the load of the first factor. In this graph, the right side denotes the tendency of smoky flavors and the left side denotes the tendency of delicious flavors. Furthermore, the whiskey brands were classified using the cluster analysis of the k-means method. To determine the number of clusters, the sum of squares of errors of prediction (SSE) within the clusters was calculated; however, no clusters with extreme decrease in SSE could be identified. Therefore, the number of clusters was set to 5 from the viewpoint of interpretability and clusters were created. Clusters 1, 2, 3, 4, and 5 contained 6, 27, 17, 6, and 30 whiskey brands, respectively. In the expanded constellation graph, each cluster was color-coded as follows. Clusters 1, 2, 3, 4, and 5 were coded as red, blue, green, purple, and yellow, respectively.

Fig. 1 shows the star at the final point in the expanded constellation graph, and Fig. 2 shows the path drawn by connecting the vectors in an order from the rightmost variable. Fig. 3 shows a constellation graph based on the mean value of each variable for each cluster.

1. From the viewpoint of the final point position, the graph results can be interpreted as follows.
   - Cluster 4 is interpreted as a cluster with a strong smoky flavor because the final point is placed on the right side of the \( \pi/2 \) direction.
   - Cluster 2 is interpreted as a cluster with a strong delicious flavor because the final point is placed on the leftmost side of each cluster.
   - Clusters 1 and 5 are interpreted as clusters with a tendency of delicious flavors.

2. From the viewpoint of the shape of the path, the graph results can be interpreted as follows.
   - Cluster 4 curves more toward the right than the other clusters.
   - In cluster 3, the star at the final point is placed near the \( \pi/2 \) direction; however, the shape of the path is curved toward the right of the other clusters. Therefore, the cluster is interpreted to have a slightly strong smoky flavor.

3. From the viewpoint of the positional relationship of the clusters, the graph results can be interpreted as follows.
   - Because the final point of cluster 5 is placed at the lowest position, it is shown that the sum of flavors is the smallest. In other words, cluster 5 is shown to have the lightest flavor.
   - However, because the final point of cluster 1 placed at the highest position, it is shown that cluster 1 has the most rich flavor.

4. From the viewpoint of the vector length, the graph results can be interpreted as follows.
   - Clusters 1 and 4 exhibit a full-body vector length, clusters 2 and 3 exhibit a medium-body vector length, and cluster 5 exhibits a light-body vector length.
   - The variable of sweetness is large in clusters 2, 5, and 1.
   - The variable of winey is large in cluster 1.
   - The variables of medicinal and smoky are large in clusters 4 and 3.
   - The variable of floral is large in clusters 2 and 5.
   - The variable of honey is large in clusters 1 and 2.

Thus, the expanded constellation graph easily shows the characteristics of whiskey brands represented by the stars and paths. Fig. 4 shows the existing constellation graph to compare with the expanded constellation graph.

The final points of objects that are close in terms of mean values and variations are placed close together. Therefore, it is difficult to grasp the difference between clusters through the position of the star. In addition, it is difficult to grasp the characteristics of the cluster through the path.

As mentioned above, the existing constellation graphs are effective for classifying objects using mean values and variations. On the other hand, the expanded constellation graphs are effective for understanding the characteristics of variables using the stars.

(3) Interpretation using the expanded kanji graph

First, a kanji was assigned to each of the nine variables, namely, medicinal = “薬,” smoky = “煙,” body = “深,” spicy = “辛,” winey = “萄,” honey = “蜜,” sweetness = “甘,” floral = “花,” and tobacco = “草.”

Table 2 shows the mean and standard deviation of the flavors for each cluster. The mean values shown in Table 2 were assigned to the height and width of the kanji. To draw the expanded kanji graph, the kanji was placed on the radar chart using the standard deviation corresponding to each variable. The value of \( \ell \) was set to 20, and the font size of the mean value \( \times 20 \) was assigned to the height and width of the kanji. Moreover, because the maximum standard deviation in all clusters was .943, the maximum value in the radar chart was set to 1. These values show that the larger the size of the kanji, the larger is the mean value of the variables. Moreover, the closer the plot of the radar chart is to the center, the smaller the standard deviation within the cluster.

Each cluster was evaluated using the expanded constellation and expanded kanji graphs as described below.
Cluster 1 is interpreted as a “full-body and winey type” because in the expanded constellation graph, the last stars of many whiskey brands belonging to cluster 1 are placed near the circumference of the semicircle and they exhibit long body vectors. The winey vector is also long, and Fig. 5 shows that whiskey brands with a deep and winey flavor are included because the “深” and “葡” are large.

Cluster 2 is interpreted as a “sweet and balanced type” because in the expanded constellation graph, the star at the final point of many whiskey brands belonging to cluster 2 is placed on the leftmost side compared with other clusters. Moreover, although the vectors of the variables, such as sweetness, honey, body, and floral, are relatively long, few whiskey brands exhibit extremely long vectors compared with those of cluster 1. Fig. 6 shows that the sizes of the kanjis other than “藥” and “草” are almost the same.

Cluster 3 is interpreted as a “smoky and balanced type” because in the expanded constellation graph, many whiskey brands belonging to cluster 3 do not have extremely long vectors and the smoky vector is longer in this cluster than in cluster 2. Additionally, the shape of the path bulges slightly to the right. Fig. 7 shows no extremely strong flavors; however “煙” is the largest. Overall, the standard deviation is small.

Cluster 4 is interpreted as a “full-body and smoky type” because in the expanded constellation graph, the star at the final point of the whiskey belonging to cluster 4 is placed on the rightmost position and close to the circumference of the semicircle. Moreover, the vectors other than medicine, smoky, and body are generally short and the flavor is heavily biased toward smoky. The expanded kanji graph in Fig. 8 shows that “煙” and “深” are the special features of cluster 4.

Cluster 5 is interpreted as a “light-body and sweet floral type” because in the expanded constellation graph, the star at the final point of the whiskey belonging to cluster 5 is placed close to the center. Further, the path of this cluster is generally shorter than those of other clusters. Fig. 9 shows that the kanji is small overall and “甘” and “花” are large.

These results reveal that the expanded kanji graph complements the expanded constellation graph and enables easy and detailed interpretation of the results.

### Table 2 Mean and standard deviation of the variables.

|       | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|-------|-----------|-----------|-----------|-----------|-----------|
| Mean  | .000      | .345      | .367      | 1.667     | 2.333     |
| SD    | .000      | .074      | .074      | .074      | .074      |

Fig.1 Expanded constellation graph based on factor 1 (only final star).
Fig. 2 Expanded constellation graph based on factor 1 (with path).

Fig. 3 Expanded constellation graph based on factor 1 (mean of each cluster).

Fig. 4 Existing constellation graph.
4. CONCLUSION

Herein, the expanded constellation and expanded kanji graphs were proposed and their usefulness was verified by evaluating the flavor data of whiskey. The expanded constellation graph has the following features.

1) The size of each variable can be grasped from the length of the vector.
2) The characteristics of the object based on the factor axis can be intuitively grasped from the shape of the path and the final position of the star.
3) It can assist in the interpretation of clusters in cluster analysis.
4) An advantage of this graph is that the overall
tendency based on the factor loading can be grasped. Moreover, the characteristics of individual variables can be visualized in one graph.

In this study, the directions of the vectors were decided by the values of the factor loadings; however it is also possible to decide the directions at equal intervals in the order of the factor loadings.

In the expanded constellation graph, the order of the variables affects the results, and deciding the order of the variables will be our future study.

The expanded kanji graph has the following features.
1) The size of the variables can be intuitively grasped from the size of the kanjis.
2) The standard deviation of the variables in each cluster can be grasped from the radar chart.
3) It becomes possible to intuitively grasp the feature of the variable by assigning a single kanji that represents the features to each variable.

The following issues must be examined in the future. First, the expanded constellation graph has the problem that only one factor can be reflected on one graph. Although the characteristics of each variable can be obtained from the shape of the path and the length of the vector, these characteristics are the result of only the first factor. The interpretation of the results in the case of multiple valid factors must be considered. Next, the expanded kanji graph is an effective method for grasping the characteristics of clusters; however, because the impression differs depending on the number of strokes and shape of the kanji, it is difficult to make a detailed comparison depending on the size of the kanji. Assigning the type of kanji is an issue for the future.

Herein, the flavor data of whiskey were used for method verification; however, we would like to verify the effectiveness of the proposed methods for other data in various fields in the future.

ACKNOWLEDGMENT: We would like to express our deepest appreciation to Prof. Mamoru Fukumori (Chugoku junior college) for giving a useful advice.

REFERENCES
1) Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, Vol. 68, pp. 361-368, 1973.
2) Andrews, D. F.: Plots of high dimensional data. Biometrics, Vol. 28, pp. 125-136, 1972.
3) Wakimoto, K.: Tree graph method for visual representation of multi-dimensional data. Journal of the Japan Statistical Society, Vol. 7, pp. 27-34, 1977.
4) Wakimoto, K. & Taguri, M.: Constellation graphical method for representing multi dimensional data. Annals of the Institute of Statistical Mathematics, Vol. 30, Pt. A, pp. 77-84, 1978.
5) Hirai, Y., Fukumori, M. & Wakimoto, K.: Kanji graph representation for multivariate data and its application to cluster analysis. Bulletin of The Computational Statistics of Japan, Vol. 1, pp. 11-21, 1988.(In Japanese).

(Received: January 8, 2021; Accepted: April 14, 2021)