Journal of Statistical
Theory & Methods
1996, Vol. 7 pp 219~226

On the Categorical Variable Clustering

Daehak Kim

Abstract Basic objective in cluster analysis is to discover natural groupings of items or variables. In general, variable clustering was conducted based on some similarity measures between variables which have binary characteristics. We propose a variable clustering method when variables have more categories ordered in some sense. We also consider some measures of association as a similarity between variables. Numerical example is included.

Keywords : cluster analysis, categorical variable, measures of association, hierarchical clustering, dendrogram

1. Introduction

Clustering is the grouping method of similar objects. Grouping is done on the basis of similarities or dissimilarities between objects or variables. Cluster analysis has been used for decades in the areas of taxonomy, medicine, anthropology, marketing research and so on. Cluster analysis is highly empirical. Different methods can lead to very different grouping, both in number of cluster and content.

There are two key steps in applying the clustering procedure. First, we need to decide on a measure of inter-object similarity. Secondly, we must specify a procedure for forming the clusters based on the chosen measure of similarity. Most efforts to produce a rather simple group structure from a complex data set necessarily require a measure of closeness or similarity. Important consideration include the nature of variables or scales of measurement and subject matter knowledge. When items are clustered, proximity is usually indicated by some sort of distance.

On the other hand, variables are usually grouped on the basis of correlation coefficient or like measures of association. In some applications, it is the variable rather than the items that must be grouped. Similarity measure for binary variables can be easily defined and widely used.

However, in reality variables can have more than two categories. So we consider variable clustering for these cases. When variables have more categories than two, we assume these categories are ordered in some sense.

Hierarchical clustering techniques proceed by either a series of successive mergers or

This research was supported by general grant, Catholic University of Taegu-Hyosung, 1996.
Daehak Kim is a professor, Department of statistics, Catholic University of Taegu-Hyosung, Kyungsan, 712-702, Korea

a series of successive divisions. Agglomerative hierarchical methods start with the individual objects. So there are initially as many clusters as objects. Eventually as the similarity decreases, all subgroups are fused into a single cluster. Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are far from the objects in the other. The results of both agglomerative and divisive hierarchical methods may be displayed in the form of two dimensional diagram known as dendrogram. Dillon and Goldstein(1984) have discussed another approach to cluster analysis, graphical methods.

Linkage methods are one of the most widely used agglomerative hierarchical method. Single linkage method merges two cluster based on smallest distance or nearest neighbor while complete linkage method are based on maximum distance. Average linkage method merges two clusters based on average distance.

In this paper, we consider variable clustering with hierarchical linkage methods when variables have more categories which is ordered in some sense. We also considered some measure of association as an similarity measure between categorical variables. Numerical example shows that proposed method works well.

2. Similarity measures for categorical variables

When the variables are binary, the original data can be rearranged in the form of a 2×2 contingency table with corresponding variables. For each pair of variables, there are n items. With the usual 0 and 1 coding, the contingency table for variable i and variable k becomes as follows.

Table 1. Contingency table

$i \backslash k$	1	0	Totals
1	a	b	$a+b$
0	c	d	$c+d$
Totals	$a+c$	$b+d$	n

The usual product moment correlation formula applied to the binary variables in the contingency table gives

$$g = \frac{ad-bc}{[(a+b)(c+d)(a+c)(a+d)]^{1/2}} \tag{1}$$

This number can be taken as a measure of similarity between two binary variables. Variety similarity measures for the table 1 were developed and adequately discussed in Johnson and Wichern(1992). Simplest ones are

$$g = \frac{a+d}{n} \tag{2}$$

and

$$g = \frac{2(a+d)}{2(a+d)+b+c}. \tag{3}$$

Where (2) gives equal weight for 1-1 matches and 0-0 matches while (3) gives double weight for 1-1 matches and 0-0 matches. Monotonicity is important because some clustering procedures are not affected if the definition of similarity is unchanged in a manner that leaves the relative orderings of similarity unchanged.

But in many practical situations, variables can be classified to many categories according to the intrinsic and relevant order. In these cases, for example, data can be rearranged as table 2 when variable i have a categories and variable k have b categories.

Table 2. $a \times b$ contingency table

$i \backslash k$	1	2	$\cdots$	b	Totals
1	N_{11}	N_{12}	$\cdots$	N_{1b}	N_{1+}
2	N_{21}	N_{22}	$\cdots$	N_{2b}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	N_{a1}	N_{a2}	$\cdots$	N_{ab}	N_{a+}
Totals	N_{+1}	N_{+2}	$\cdots$	N_{+b}	n

For categorical variables, agreement of a single variable with a partition is expressed by a contingency table in which one margin is the variable and the other margin is the partition. The partition is nothing but a category variable itself which is an amalgamation of all the category variables in the data, a category version of the first principal component for continuous variables. Between two partitions, the above measures of agreement are undesirable.

There are many kinds of measures of association between categorical variables. Traditional measure of association is the chi-square statistic

$$\chi^2 = n(\sum_i \sum_j \frac{N_{ij}^2}{N_{i+} \cdot N_{+j}} - 1). \tag{4}$$

Another measures of association for the table 2 suggested by Goodman and Kruskal(1979) include

$$\lambda_b = (\sum_i N_{im} - N_{+m}) / (1 - N_{+m}) \tag{5}$$

and

$$\lambda_a = (\sum_j N_{mj} - N_{m+}) / (1 - N_{m+}), \tag{6}$$

where

$$N_{+m} = \max_j N_{+j}, \quad N_{m+} = \max_i N_{i+}, \quad N_{im} = \max_{1\le j\le n} N_{ij}, \quad N_{mj} = \max_{1\le i\le n} N_{ij} \tag{7}$$

But these measure of associations can't directly be used as a similarity measure since we are assuming the categories are ordered in some sense. So we consider another measures of association which is appropriate for ordered categories. One possible method suggested by Hartigan(1975) is to use

$$g = \Sigma_i \Sigma_j N_{ij} \log N_{ij} - \Sigma_i N_{i+} \log N_{i+} - \Sigma_j N_{+j} \log N_{+j} + nN_{ij} \tag{8}$$

which is called information measure. This is just the log likelihood ratio of the general multinomial hypothesis against the hypothesis of independence between variables. But it is not proper for our cases. We consider another measure of association G propsed by Goodman and Kruskal(1972)

$$G = \frac{P_s - P_d}{1 - P_t} = \frac{2P_s - 1 + P_t}{1 - P_t} \tag{9}$$

where

$$P_s = \frac{2}{n^2} \Sigma_i \Sigma_j N_{ij} \{\sum_{i'>i} \sum_{j'>j} N_{i'j'}\}, P_d = \frac{2}{n^2} \Sigma_i \Sigma_j N_{ij} \{\sum_{i'>i} \sum_{j'<j} N_{i'j'}\}$$

and

$$P_t = n^{-2} \Sigma_i \Sigma_j N_{ij} \{N_{i+} + N_{+j} - N_{ij}\} = n^{-2} \{\Sigma_i N_{i+}^2 + \Sigma_j N_{+j}^2 - \Sigma_i \Sigma_j N_{ij}^2\}.$$

It must be noted that G tells us how much more probable it is to get like than unlike orders in two classifications.

[Remark] Some important properties of G is as follows. G is indeterminate if all count is entirely in a single row or column of the table. G is 1 if all count is concentrated in an upper left to lower right diagonal of the table. G is -1 if all count is concentrated in a lower left to upper right diagonal of the table. G is 0 in the case of independence, but the converse need not hold except in the 2×2 case.

[Lemma] For the efficient calculation of G for the table 2, let's assume the following notations

$$S_{ij} = \sum_{i'>i} \sum_{j'>j} N_{i'j'} + \sum_{i'<i} \sum_{j'<j} N_{i'j'}, \quad D_{ij} = \sum_{i'<i} \sum_{j'>j} N_{i'j'} + \sum_{i'>i} \sum_{j'<j} N_{i'j'}$$

$$P_{ij}^s = N_{ij} \times S_{ij}, \quad P_{ij}^d = N_{ij} \times D_{ij}, \quad P_s = \sum_i \sum_j P_{ij}^s, \quad P_d = \sum_i \sum_j P_{ij}^d$$

Then G can be written as follows which is simple function of P_s and P_d

$$G = \frac{P_s - P_d}{P_s + P_d} \tag{10}$$

3. Numerical studies

As explained in section 1, after some similarity measures are chosen, for the clustering of variable, we should consider the procedure for forming the clusters. In this paper, we use the most popular one, agglomerative hierarchical method, simple linkage, complete linkage and average linkage respectively. We didn't consider non hierarchical method, for example, the k means method because it is designed for the observation clustering rather than variable clustering and the number of clusters are needed.

Table 3. Original data

obs	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
15	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Table 4. Transformed data

obs	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
1	2	1	2	1	2	2	1	1
2	2	1	2	3	2	2	1	2
3	3	4	1	1	2	2	1	2
4	2	2	2	2	1	1	2	1
5	3	1	3	1	1	1	1	4
6	3	3	1	2	1	3	2	2
7	2	3	3	4	2	2	1	3
8	2	1	4	2	2	3	1	1
9	3	3	2	3	4	2	1	2
10	2	3	3	1	2	1	3	1
11	1	1	3	1	3	4	1	1
12	2	2	3	3	2	1	1	4
13	2	3	3	1	3	2	3	1
14	2	2	1	1	2	2	1	1
15	1	1	2	3	1	1	1	3
16	2	1	4	2	4	4	1	1
17	1	1	2	3	4	1	1	4
18	2	3	2	2	2	3	1	2
19	2	2	1	1	1	4	1	1
20	2	2	2	2	2	2	3	1
21	2	1	3	3	2	2	1	4
22	2	1	3	1	3	3	2	3

We considered the public utility company data as an numerical example. The data are found in Johnson and Wichern (1992, p593). It consists of 22 observations and 8 variables and appeared in table 3. Variable X_1 through X_8 represents fixed charge coverage ratio, rate of return on capital, cost per KW capacity in place, annual load

factor, peak KWH demand growth, sales, percent nuclear and total fuel costs respectively.

For the categorical variable clustering of these continuous data, we transformed the data appropriately. For the transformation, we considered a standard deviation of each variable in order to get reasonable the category number. We redefined the 8 variable as

$$Y_1=\begin{cases}1 & \text{if } X_1 \le 1. \\ 2 & \text{if } 1 < X_1 \le 1.25 \\ 3 & \text{if } X_1 > 1.25\end{cases} \qquad Y_2=\begin{cases}1 & \text{if } X_2 \le 10 \\ 2 & \text{if } 10 < X_2 \le 12 \\ 3 & \text{if } 12 < X_2 \le 14 \\ 4 & \text{if } X_2 > 14\end{cases}$$

$$Y_3=\begin{cases}1 & \text{if } X_3 \le 130 \\ 2 & \text{if } 130 < X_3 \le 170 \\ 3 & \text{if } 170 < X_3 \le 210 \\ 4 & \text{if } X_3 > 210\end{cases} \qquad Y_2=\begin{cases}1 & \text{if } X_2 \le 10 \\ 2 & \text{if } 10 < X_2 \le 12 \\ 3 & \text{if } 12 < X_2 \le 14 \\ 4 & \text{if } X_2 > 14\end{cases}$$

$$Y_5=\begin{cases}1 & \text{if } X_5 \le 1 \\ 2 & \text{if } 1 < X_5 \le 4 \\ 3 & \text{if } 4 < X_5 \le 7 \\ 4 & \text{if } X_5 > 7\end{cases} \qquad Y_6=\begin{cases}1 & \text{if } X_6 \le 6500 \\ 2 & \text{if } 6500 < X_6 \le 10000 \\ 3 & \text{if } 10000 < X_6 \le 13500 \\ 4 & \text{if } X_6 > 13500\end{cases}$$

$$Y_7=\begin{cases}1 & \text{if } X_7 \le 18 \\ 2 & \text{if } 18 < X_7 \le 36 \\ 3 & \text{if } X_7 > 36\end{cases} \quad \text{and} \quad Y_8=\begin{cases}1 & \text{if } X_8 \le 0.8 \\ 2 & \text{if } 0.8 < X_8 \le 1.3 \\ 3 & \text{if } 1.3 < X_8 \le 1.8 \\ 4 & \text{if } X_8 > 1.8\end{cases}$$

Based on these transformations, we can get categorized data and the results are represented in table 4. For the data in table 3, the inputs for variable cluster analysis was correlation matrix and this is represented as follows.

$$R=\begin{bmatrix}
1.000 & & & & & & & \\
.643 & 1.000 & & & & & & \\
-.103 & -.348 & 1.000 & & & & & \\
-.082 & -.086 & .100 & 1.000 & & & & \\
-.259 & -.260 & .435 & .034 & 1.000 & & & \\
-.152 & -.010 & .028 & -.288 & .176 & 1.000 & & \\
.045 & .211 & .115 & -.164 & -.019 & -.374 & 1.000 & \\
-.013 & -.328 & .005 & .486 & -.007 & -.561 & -.185 & 1.000
\end{bmatrix}$$

Proposed measures of association were adapted to the data in table 4 and calculated values of G for each pairs of variables are represented in the following matrix S. It should be noted that the value of G between variable 7 and 8 equals to 0 while the correlation between variable 7 and 8 are $-.185$. The other values of S are similar to those of R. All computation including correlation and measures of association matrix R was carried out on workstation based on FORTRAN programs.

Figure 1 represents dendrogram of the original data where complete linkage is used

and the inputs for clustering is correlation matrix R. Figure 2 represent the dendrogram where the only different thing from figure 1 is input data of similarity matrix S of categorized data.

$$S = \begin{bmatrix} 1.000 & & & & & & & \\ .674 & 1.000 & & & & & & \\ -.083 & -.139 & 1.000 & & & & & \\ -.126 & .044 & -.136 & 1.000 & & & & \\ -.152 & -.041 & .050 & .049 & 1.000 & & & \\ .096 & -.003 & -.127 & -.221 & .126 & 1.000 & & \\ .000 & .165 & -.138 & -.016 & -.034 & -.093 & 1.000 & \\ -.038 & -.063 & .011 & .376 & -.027 & -.213 & .000 & 1.000 \end{bmatrix}$$

All of the results of cluster analysis is based on the IMSL subroutines, *clink, pgopt* and *treep* which works on workstation. The two plots show similar forms in cluster. Variable 1 and 2, variable 4 and 8 cluster at intermediate similarity levels. The final merger brings together the (12478) group and (356) group in both dendrograms.

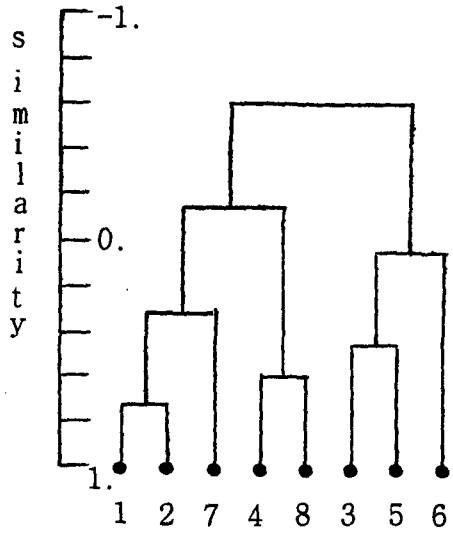

Figure 1. Dendrogram of complete linkage

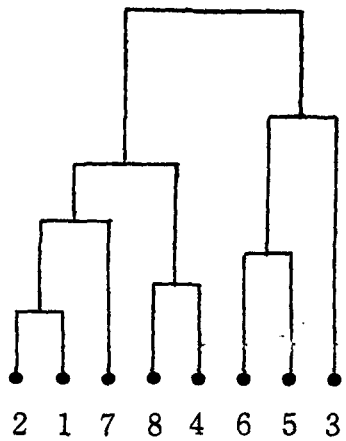

Figure 2. Dendrogram of complete linkage

Figure 3 and figure 4 represent dendrograms of the variable cluster. Simple linkage method is used in figure 3 while average linkage method is applied to figure 4. These are based on proposed measures of similarity matrix R. In this example, simple linkage brings the final two clusters (12357) and (468) while average linkage brings the two cluster (125) and (34678). These results are some what surprise. When correlations are used as similarity measures, variables with large negative correlations are regared as

very dissimilar variabls with large positive correlations are regarded as very similar. In this case, the distance between cluster is measured as the smallest between members of the corresponding clusters. Complete linkage method looks appropriate for the considered example.

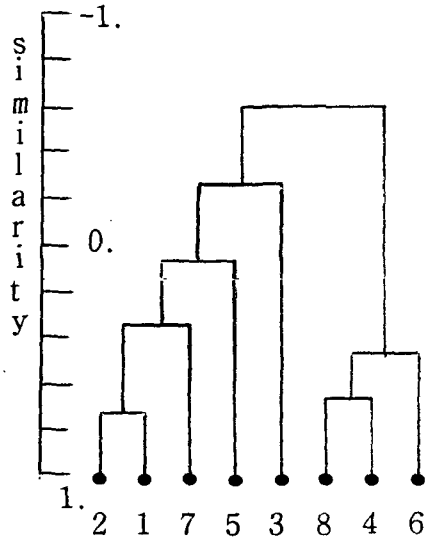

Figure 3. Dendrogram of simple linkage

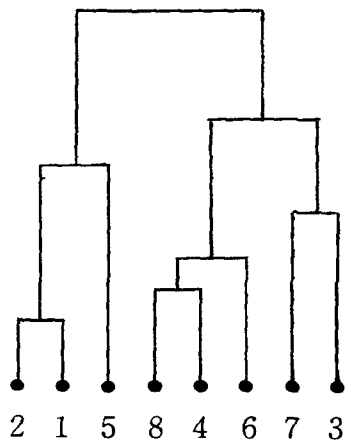

Figure 4. Dendrogram of average linkage

References

Dillon, W. R. and Goldstein, M. (1984) *Multivariate Analysis : Methods and Applications,* John Wiley & Sons.

Goodman, L. A. and Kruskal, W. H. (1979), Measures of Association for cross classifications Ⅳ: Simplification of asymptotic variance, *Journal of the American Statistical Association, 67, 415-421*

Goodman, L. A. and Kruskal, W. H. (1979), *Measures of Association for cross classifications,* Springer-verlag, New York Heidelberg, Berlin.

Hartigan, A. J.(1975), *Clustering Algorithm,* John Wiley & Sons.

IMSL(1991), User's Manual : *Stat/Library and Math/Library*

Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis,* Prentice Hall, Engelwood Cliffs, New Jersey.