Clustering Methods for Ordinal Variables

© Ferenc RUFF
Szent István University, Gödöllő, Hungary
Ruff.Ferenc@gtk.szie.hu

The first aim of this article is to present a clustering method, which can be used in the case of ordinal dependent variables. The importance of the topic lies in that e.g. in marketing research, we often have to work with the abovementioned variables. After presenting the LCA method, it has been compared with other well-known methods. The second aim is to find the best clustering method(s) among the applicable ones with the help of simulation tests.

Introduction

Many diverse procedures have been developed to discover homogeneous groups (segments), e.g. the different variants of perceptual maps, decision trees and cluster analysis. Some of these methods play a significant role in scientific examinations, and in the case of practical applications (e.g. K-means method), while others have only been used by some researchers who are expressing an interest in these algorithms. Of course, it depends on the software products, which contain these algorithms, so which are used by professionals of the scientific and economic life in their daily work.

Simon (2006) has a wide-range overview about the methods of cluster analysis and their opportunities of application. According to her opinion, it is necessary to prepare the analysis carefully by the following operations:

- the definition of the objects,
- the treatment of the outliers,
- the definition of the variables,
- the definition of the possible weighting of the variables,
- the examination of the comparability of the variables.

Of course, she deals with the problem of measurement scale of the variables. She points out that the specialists' opinion is not uniform in this issue. She outlines procedures on a practical manner onto not-metric (nominal) variables, and concerning procedures onto metric (measured on interval or ratio scale) variables. In this article, such a cluster theory has been reviewed, in which there is no restriction on data types, but inside this an algorithm has been tested that was developed for ordinal variables.

Among the cluster procedures applied in the area of marketing research the most applied is the K-means method (in the group of the non-hierarchical methods). There are some problems about this clustering algorithm, which queries the received result though:

- Mostly the Euclidean distance is applied onto the measurement of the similarity.
- Applying this distance on the ordinal scale may cause problems.
- The result depends on the starting state (that is the initial centres).
- It is necessary to give the number of the clusters in advance.
- Does not find the absolute optimum inevitably (may stop at local optimum).

The elimination of these disadvantages cannot be realised totally, therefore a less known method has been outlined as a possible alternative. The actuality of applying this method is that a part of the variables applied in the marketing research are measurable on an ordinal scale many times, so in such a case the application of Euclidean distance makes the result doubtful. This method does not group the individuals by the measure of the similarity (or distance), but with the help of a statistical model. The method can be applied in the practical marketing research, in the education, and in the scientific research as well.

Presentation of the method

The theoretical background of the method is the Latent Class Analysis (LCA), which tries to explore the invisible contexts (latent variable) behind the observed variables with statistical means (Goodman, 1974). The latent variable is a nominal variable with determined number of values.

The essence of the method is the estimation of the parameters of the population, based on the sample with statistical tools. The basis of the parameter estimation is the principle of the largest probability (maximum likelihood method, ML). It is not necessary to apply restrictive conditions to the dependent variables, and the linear transformation of the variables does not influence the result. Furthermore, it is also allowed to use variables measured on different scales, at the same time.

Let X be a latent variable and $\boldsymbol{y}_k$ be the k-th observed vector variable (e.g. answers of the k-th question), where $\mathbf{1} \leq k \leq K$. Now let the observed variables be measurable on ordinal scale (because in the field of marketing research it is commonly used, so we reduce the theory onto this case), and let D_k be the number of values that the k-th variable can takes on. Let the number of the values, which can be taken by the latent variable (i.e. the number of clusters), be C. Let $\boldsymbol{A}$ be the matrix which contains the $\boldsymbol{y}_k$vectors, and let $\boldsymbol{z}_i$ be the i-th row of this matrix (i-th object, i.e. the answers of the i-th person). According to the model, the probability that a randomly selected object (let's denote it with $\boldsymbol{Z}$) takes on a given values (e.g. $\boldsymbol{Z} = \boldsymbol{z}_i$), it is influenced by a background (not known) variable (X) (theorem of total probability):

$$\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i) = \sum_{j=1}^{C} \mathrm{P}(X = x_j)\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i | X = x_j) \tag{1}$$

where x_j is the j-th value of the latent variable.

Our question is: what is the probability of a given value of $\boldsymbol{Z}$ (e.g. $\boldsymbol{z}_i$) falling into the j-th cluster ($\mathbf{1} \leq j \leq C$) defined by the latent variable (Bayes theorem)?

$$\mathrm{P}(X = x_j | \boldsymbol{Z} = \boldsymbol{z}_i) = \frac{\mathrm{P}(X = x_j)\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i | X = x_j)}{\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i)} \tag{2}$$

To calculate the requested probabilities it is necessary to define the value of the following parameters: $\mathrm{P}(X = x_j)$ and $\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i | X = x_j)$.

Presupposing the independence of the variables ($\boldsymbol{y}_k$) this can be expressed by the help of the current values of the variable $\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i | X = x_j) = \prod_{k=1}^{K} \mathrm{P}(Z_k = y_{\mathbf{k}_i} | X = x_j)$, where Z_k is the k-th component of $\boldsymbol{Z}$, and $y_{\mathbf{k}_i}$ is the i-th component of the variable $\boldsymbol{y}_k$.

The procedure calculates these probabilities by the help of the maximum likelihood method. We are looking for the extreme value (maximum) of the logarithm

of the likelihood function [equation (3)]. The optimisation is made by the Expectation-Maximization algorithm (EM) (Dempster et al., 1977).

$$\log L = \sum_{i=1}^{N} \log \mathrm{P}(X = x_j)\mathrm{P}(\boldsymbol{Z} = \boldsymbol{z}_i | X = x_j) \quad (3)$$

The calculations have been made by the "R" software (R Development Core Team 2011), and within the R the poLCA package has been used (Linzer 2007). This algorithm was developed to examine variables with an ordinal measurement level.

Materials and methods

Four clustering methods have been involved in the examinations: K-means, LCA, Hierarchical (HC), K-medians. These algorithms are well known for marketing researchers, because these are the most applied tools (cf. clustering methods of SPSS). The K-means and HC are the most popular methods, and the K-medians was mentioned e.g. by Řezanková (2009) for ordinal variables.

The calculations have been made by the "R" software (R Core Team, 2013), and within the R some packages have been used:

- K-means: the kmeans function of the stats package has been used with Euclidian distance as a similarity measure.
- LCA: the poLCA function of the poLCA package has been used (Linzer 2007).
- HC: the hclust function of the stats package has been used. Method: "ward".
- K-medians: the kcca function of the flexclust package has been used with Manhattan distance as a similarity measure.

10 experiments have been performed, and in every experiment 30 databases have been prepared. Every database contains 1000 rows (observation units) and 4 columns (variables). The results of the 30 databases have been compared for the 4 methods: hypothesis test for the mean (to compare the result of the two methods).
Two different tests have been applied:

- paired samples t-test (if the necessary conditions have been met),
- paired samples Wilcoxon test (in other cases).

The databases are random samples from the 4-dimensional space. Every variable (there are 4 variables altogether) have been simulated as a realization of a binomial random distribution with specified parameters. Values can be recorded by the binomial random variables are follows: 0, 1, 2, 3, 4, 5, 6. So it can be an ordinal scale (often used in marketing research).

Every database contains two clusters, which have been configured by the parameters of the variables. The binomial random variable has two parameters: n and p. n = 6 in all the cases. The values of p in the first cluster are: p_1 = 0.7, p_2 = 0.9, p_3 = 0.8, p_4 = 0.8 (because there are four variables so there are four binomial random variables). The values of p in the second cluster can be seen in Table 1 (there are 10 rows according to the 10 experiments).

Table 1: The four parameters of the variables in the second cluster

	p_1	p_2	p_3	p_4
1	0.3	0.2	0.2	0.1
2	0.3	0.2	0.2	0.2
3	0.4	0.3	0.3	0.3
4	0.4	0.3	0.3	0.4
5	0.5	0.3	0.4	0.4
6	0.5	0.4	0.4	0.5
7	0.6	0.4	0.6	0.6
8	0.6	0.5	0.6	0.7
9	0.7	0.5	0.7	0.7
10	0.7	0.6	0.7	0.8

Results

As it can be seen in Table 1, during the 10 experiments the two clusters are getting closer. In this case – as it was to be expected - the numbers of correct classifications have been reduced for each method.

a) First version of the simulation: the numbers of elements of the two clusters are 1000 and 1000.

Table 2: The mean and the relative standard deviation of the numbers of correct classifications and the results of their comparisons in the 10 experiments (first version)

	KM mean (RSD%)	LCA mean (RSD%)	HC mean (RSD%)	KMED mean (RSD%)	p-val.1	p-val.2
1	2000 (0.028)	2000 (0.031)	1999 (0.046)	2000 (0.031)	0.773	0.675
2	1999 (0.06)	1999 (0.061)	1997 (0.117)	1999 (0.055)	0.386	0.667
3	1990 (0.132)	1991 (0.134)	1985 (0.242)	1988 (0.180)	0.009	0.000
4	1987 (0.170)	1989 (0.174)	1981 (0.334)	1982 (0.170)	0.003	0.000
5	1976 (0.237)	1980 (0.252)	1970 (0.395)	1970 (0.444)	0.000	0.000
6	1950 (0.392)	1954 (0.381)	1926 (0.870)	1932 (0.581)	0.000	0.000
7	1886 (0.570)	1897 (0.580)	1833 (1.842)	1822 (1.867)	0.000	0.000
8	1796 (0.587)	1818 (0.744)	1734 (2.251)	1680 (10.820)	0.000	0.000
9	1746 (0.964)	1751 (2.336)	1666 (4.083)	1569 (10.078)	0.020	0.000
10	1644 (1.355)	1487 (7.877)	1533 (4.598)	1431 (11.129)	0.000	0.241

p-val.1: comparison of the means of LCA and KM
p-val.2: comparison of the means of LCA and KMED

In all the position of the two clusters, the classification accuracy is better in the case of K-means and LCA as in the case of HC and K-medians. Of course, the difference is higher if the clusters are closer to each other. Therefore, the hypothesis

tests have been performed in such cases: LCA vs. K-means and LCA vs. K-medians. 7 times in the case of the 10 experiments, the average of the LCA is higher than the average of the K-means (at 5% significance level). The result is the same even for the LCA and K-medians. Namely in these cases the differences of the means are statistically significant.

There is a PhD thesis (Anderlucci, 2012), in which two clustering methods (Latent class clustering (LCC) and Partitioning around medoids (PAM)) were examined in many simulation experiments. Their conclusion was that most of the experiments ended with better performance of the LCC, mainly when the clusters were closely to each other. In the case of the other experiments, the accuracy of the two methods was almost the same. These results have been confirmed in this article, too. However, there is a further result: the performance of the K-means algorithm is almost the same as the performance of the LCA.

b) Second version of the simulations: the numbers of elements of the two clusters are 500 and 1500. In this case, the difference between the mean of LCA and KM is not statistically justified (Table 3). However, 8 times the average of the LCA is higher than the average of the K-medians. The difference is getting bigger by reducing the distance between the two clusters. The result is similar as in the previous experiment: the best methods are LCA and K-means.

Table 3: The mean and the relative standard deviation of the numbers of correct classifications and the results of their comparisons in the 10 experiments (second version)

	KM mean (RSD%)	LCA mean (RSD%)	HC mean (RSD%)	KMED mean (RSD%)	p-val. 1	p-val. 2
1	1999 (0.039)	1999 (0.047)	1999 (0.096)	1999 (0.033)	0.080	0.120
2	1999 (0.058)	1999 (0.069)	1998 (0.087)	1999 (0.080)	0.212	0.374
3	1993 (0.133)	1993 (0.119)	1989 (0.201)	1987 (0.186)	0.062	0.000
4	1991 (0.140)	1990 (0.115)	1984 (0.373)	1983 (0.216)	0.120	0.000
5	1984 (0.235)	1983 (0.196)	1972 (0.607)	1976 (0.486)	0.020	0.000
6	1963 (0.391)	1962 (0.381)	1942 (0.873)	1906 (7.606)	0.133	0.000
7	1919 (0.551)	1915 (0.567)	1876 (1.079)	1714 (14.304)	0.001	0.000
8	1850 (0.631)	1852 (0.692)	1791 (1.595)	1538 (18.89)	0.325	0.000
9	1817 (0.753)	1808 (1.673)	1766 (1.858)	1332 (18.287)	0.130	0.000
10	1708 (0.747)	1626 (7.491)	1596 (10.067)	1270 (16.132)	0.001	0.000

p-value1: comparison of the means of LCA and KM
p-value2: comparison of the means of LCA and KMED

Conclusions

The aim of this article has been to compare different clustering methods on such databases that contain only ordinal variables. The performance of the most popular K-means clustering method is almost the same as the outcomes of the LCA method. It is remarkable because the K-means uses of the Euclidian distance as a similarity measure, and - however - the LCA was developed for ordinal variables. The accuracy of the other two methods is lower, compared with the previous ones. It is difficult to select the proper method - e.g. for clustering a database – because of the many quantitative methods. Of course, there are several "new" methods, which have not yet become known, but their performance may be better than the "old" ones. In this article, a slightly known method has been presented in addition to the better-known ones. The simulations show the utility of the LCA method e.g. in the marketing research.

References

ANDERLUCCI, L. (2012). Comparing Different Approaches for Clustering Categorical Data. PhD thesis. Alma Mater Studiorum - Università di Bologna.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (With Discussion). *Journal of the Royal Statistical Society*, 39, Series B, 1-38.

GOODMAN, L. (1974). The analysis of systems of qualitative variables. *American Journal of Sociology,* 79, 1179-1259.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [http://www.R-project.org/]

ŘEZANKOVÁ, H. (2009). Cluster analysis and categorical data. *Statistika*, 216-232.

SIMON J. (2006). A klaszterelemzés alkalmazási lehetőségei a marketingkutatásban. *Statisztikai Szemle,* (7), 627-650.