STOWARZYSZENIE EKONOMISTÓW ROLNICTWA I AGROBIZNESU
Roczniki Naukowe • tom XIV • zeszyt 6

Ferenc Ruff

Szent István University, Gödöllő, Hungary

EMPIRICAL COMPARISON OF A MODEL BASED AND A NON MODEL BASED CLUSTERING METHODS

PORÓWNANIE EMPIRYCZNE MODELOWYCH I NIEMODELOWYCH METOD GRUPOWANIA

Key words: marketing research, quantitative methods, cluster analysis, model based method

Słowa kluczowe: badania marketingowe, metody ilościowe, analizy grupowania, metody bazujące na modelach

Abstract. The aim of the examination is to draw the attention to the usage of procedures in the field of clustering, which can handle ordinal variables without distance measurement (e.g. Euclidean distance) and leads to a significantly more accurate result furthermore.

Introduction

For a long time cluster analysis has a great importance in marketing research, and it is the most frequently applied method in the practical field of market segmentation and cognition of markets. Thanks to the softwares implying these methods there are widespread algorithms [Wedel, Kamakura 2000]. These are primarily the non model-based procedures: e.g. K-means method, or the Ward method from the hierarchical algorithms. The role of model-based procedures is limited even in scientific works, as well as in the applications in practice [Andrews et. al. 2010]. However, it is worth to deal with these methods as the type and structure of the data collected in marketing research.

Among these variables we can find variables measured on an ordinal scale and others can be measured on a numerical one (e.g. opinion on a given product). To take into account variables like these in the same examination requires due foresight on behalf of the researcher while choosing the applied method.

Materials and methods

The aim of the present article is to make an empirical comparison of a non model based method – the Twostep Cluster which is a method available in SPSS [The SPSS… 2001] , and a model based method – Latent Class Analysis (LCA) [Goodman 1974]. Both methods are suitable for the treatment of variables with an ordinal measurement level, so they are suitable for marketing research (e.g. examination of the market segmentation). The methods have been applied by the help of SPSS PASW Statistics 18 software package and the R environment [R Development … 2011].

These algorithms may provide easily useable scientific methods (which can be easily interpreted) either in the education and in the academic research or for the firms working in the competitive sector.

The examinations have been carried out on databases generated randomly. The constructions of the databases are equal to a table coming from a questionnaire survey, in which the number of the variables (NOV) are 3 and 5, the number of objects (NOO) are 1000 and 5000 (Tab. 1). The number of clusters (NOS) has been set to 2 and 4. Onto each single variant 5 different databases have been made, i.e. there are 40 pieces altogether (2 x 2 x 2 x 5). The values of the variables have been resulted by random

Table 1. Parameters of the databases involved in the examinations
Tabela 1. Parametry baz danych wykorzystywanych w badaniach

Number of clusters/ ***Liczba klastrow NOS***	**Number of variables/** ***Liczba zmiennych NOV***	**Number of objects/** ***Liczba obiektów NOO***
2	5	1000
2	3	1000
2	3	5000
2	5	5000
4	5	5000
4	3	5000
4	5	1000
4	3	1000

Source: own study
Źródło: opracowanie własne

binomial distribution (with different parameters). These records are integer values which can vary from 1 to 5, which, in marketing research, is the commonly applied Lickert scale (it also means ordinal scale). The number of the clusters – which can be found in the databases – have been set up for the parameters of the variables (in case of a binomial distribution these are *n* and *p*, but in these cases *n* is a constant (*n*=4).

Methods involved in the examination

Since the two methods mentioned above are not widely known [Bacher et. al. 2004], a short review of these algorithms is given in this part of the article.

Twostep Cluster [Chiu et. al. 2001]. The method was developed onto the analysis of the databases implying many objects in such a way that should be suitable for the processing of variables with different measurement level. It makes the formation of the clusters in two stages:

1. The pre-processing step, in which all objects – looking at all the database throughout – are ordered into groups. The algorithm builds a tree, in which each object has a place. The objects which are in the same node, will be considered as a new single object henceforth. The information characterizes these groups (as objects) in the additional stages are follows: the number of elements in the group, the mean and the variance of the individuals forming the group [Zhang et al. 1996]. With the help of this data-compression, the algorithm is suitable for processing very big databases. The formed groups can be observed where the pattern condenses. The software applies different distance measures for different type of data [The SPSS... 2001].
2. The number of the groups received is much smaller, than the number of the original objects. The grouping of these new objects takes place in the second phase. For this purpose, the SPSS software uses the agglomerative hierarchical clustering method.For finding the number of clusters, an automatic procedure has been drawn up, that consist of the Bayesian Information Criterion (BIC) on the one hand and the analysis of the distance between the clusters on the other hand.

Latent Class Analysis. The theoretical background of the method is trying to explore the non visible relationship (latent variable) behind the observed variables with statistical devices [Goodman 1974]. The latent variable is a nominal variable with a particular number of values. The number of values of the latent variable has to be given in advance. The essence of the method is the estimation of the parameters of the population by the help of the sample by statistical process. The basis of the parameter estimation is the principle of the largest probability (maximum likelihood method, ML). Let *X* be the latent variable, and Y_k the *k*-th observed variable ($1 \leq k \leq$ K). Let the observed variable be measured on ordinal scale, and the number of the values of the *k*-th variable be D_k. Let the number of clusters (the number of values of the latent variable) be *C*. Let *Y* be a matrix which contains the vectors Y_k (object – attribute matrix), and y_i is the *i*-th row of this matrix (*i*-th object). The probability of *Y* taking a certain value (e.g. $Y = y_i$) is determined by an unknown "latent" variable (*X*), as it is shown by the Law of total probability in eq. (1):

$$\mathbf{P}(\boldsymbol{Y} = \boldsymbol{y}_i) = \sum_{x=1}^{C} \mathbf{P}(X = x)\mathbf{P}(\boldsymbol{Y} = \boldsymbol{y}_i | X = x) \tag{1}$$

The question: what is the probability of a certain object (y_i) belongs to the *x*-th cluster. See the Bayes theorem in eq. (2):

$$\mathbf{P}(X = x | \boldsymbol{Y} = \boldsymbol{y}_i) = \frac{\mathbf{P}(X = x)\mathbf{P}(\boldsymbol{Y} = \boldsymbol{y}_i | X = x)}{\mathbf{P}(\boldsymbol{Y} = \boldsymbol{y}_i)} \tag{2}$$

We can calculate this probability if we know the value of parameters as follows: *P*($X = x$) and P($Y = y_i | X = x$). To take into consideration the independency of the variables:

P($Y = y_i | X = x$) $\prod_{k=1}^{k} \mathrm{P}(Y_k = y_{ik} | X = x)$. The LCA method calculates these parameters by the help of Maximum Likelihood method. The method finds the maximum of the logarithm of the likelihood function[1] (eq. (3). The theory of the ML method is finding the parameters of the population

$$\log \mathbf{L} = \sum_{i=1}^{N} \log \sum_{x=1}^{C} \mathbf{P}(X = x)\mathbf{P}(\boldsymbol{Y} = \boldsymbol{y}_i | X = x) \tag{3}$$

[1] By the help of Expectation-Maximization (EM) algorithm.

The execution of the calculations has been made with the poLCA package of the R environment [Linzer 2007]. This algorithm can handle only ordinal variables (this is sufficient for this examination). Applying e.g. the Euclidean distance in case of variables like these would be a conceptual error (e.g. the K means method doesn't appropriate in this case).

Results

The number of clusters (NOS) was known through the preparation of the databases. It has been utilized during the running of the algorithms (MOD)[2]: in case of Twostep method the maximum number of clusters (input data of the algorithm) and in case of poLCA the number of the clusters (input data of the algorithm) has been adjusted based on NOS. In this case the ability of poLCA method to determine the number of clusters has got out from the examination, although in the case of Twostep an upper limit was also introduced.

The accuracy of the models was compared by the help of Adjusted Rand Index (ARI) [Rand 1971, Hubert and Arabie 1985]. The values of the ARI fall into [0;1] interval – the bigger the value, the more accurate the model is. All the models have given a number by the comparison of the estimated classifications (constructed by the algorithm) and the valid classifications. After comparing the results one can see that the model based procedure (poLCA) received a better value in each case, so created a more accurate categorisation, than the non model based (Twostep) procedure. For examination of the diversity of the ARI values one-way analysis of variance has been used (Tab. 2). Naturally, it confirmed that the average values of the accuracy of the models created by the two procedures differ from each other significantly.

Table 2. The table of one-way ANOVA
Tabela 2. Model jednokierunkowy ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MOD	1	0,9085	0,9085	38,01	0,0000
Residuals/ *Pozostałe*	78	1,8643	0,0239		

Source: own study
Źródło: opracowanie własne

Because of comparing the accuracy with other variables multiway analysis of variance was executed (Tab. 3). From the four variables three proved to be significant:

– the model (MOD) variable (already examined before),
– the number of clusters (segments) (NOS),
– the number of the variables (NOV).

From the examination we can also get an answer to questions like this: the interaction of two variables has a considerable effect on the values of the target variable.

We can see (Table 3) that two couples proved to be significant watching of their effects simultaneously:

– the number of the objects (NOO) and the number of the variables (NOV),
– the number of clusters (NOS) and the number of the variables (NOV).

We can observe the direction of their relationship on figure 1 and on figure 2. On the vertical axis the average of the ARI indicator is shown and the confidence intervals belonging to him (on 0.95 confident level). Figure 1 shows that in the case of the greater number of variables (NOV=5) we receive a more inaccurate model with the greater number of observation (NOO=5000), while in the case of smaller number of variables (NOV=3) it is the contrary.

Table 3. The table of multiway ANOVA
Tabela 3. Model wielokierunkowy ANOVA

Variable/ ***Zmienna***	**Sum Sq**	**Df**	**F value**	**Pr(>F)**	**Eta-sq**	**P. Eta-sq**	**Omega-sq**
MOD	0.90847	1	156.2066	0.00000	0.32764	0.70936	0.32486
NOO	0.01902	1	3.2702	0.07525	0.00686	0.04861	0.00475
NOS	0.18670	1	32.1020	0.00000	0.06733	0.33404	0.06510
NOV	0.80390	1	138.2277	0.00000	0.28992	0.68352	0.28722
MOD:NOO	0.01123	1	1.9312	0.16944	0.00405	0.02929	0.00195
MOD:NOS	0.01605	1	2.7595	0.10157	0.00579	0.04134	0.00368
NOO:NOS	0.00000	1	0.0003	0.98730	0.00000	0.00000	0.00000
MOD:NOV	0.00229	1	0.3939	0.53249	0.00083	0.00612	0.00000
NOO:NOV	0.07374	1	12.6800	0.00070	0.02659	0.16536	0.02445
NOS:NOV	0.27861	1	47.9057	0.00000	0.10048	0.42809	0.09818
…							
Residuals/ *Pozostałe*	0.37221	64					

Source: own study
Źródło: opracowanie własne

[2] Model: Twostep or poLCA

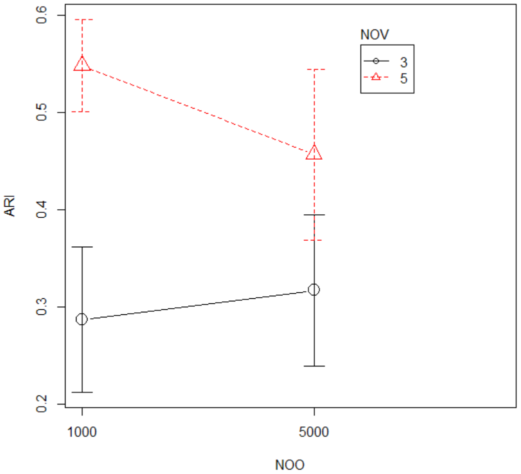

Figure 1. The relationship of the averages (NOO-NOV)
Rysunek 1. Zależności pomiędzy średnimi (NOO-NOV)
Source: own study
Źródło: opracowanie własne

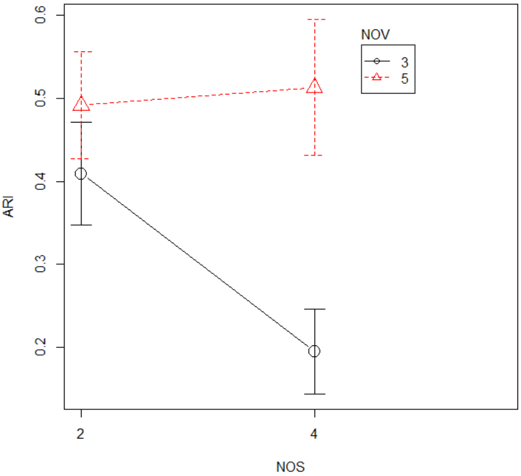

Figure 2. The relationship of averages (NOS-NOV)
Rysunek 2. Zależności pomiędzy średnimi (NOS-NOV)
Source: own study
Źródło: opracowanie własne

Of course it is necessary to handle this result circumspectly, since the distributions of the variables may influence these associations (but this article doesn't examine this problem). Figure 2 shows that in the case of smaller number of variables (NOV=3) the increase of the number of clusters worsened the average ARI value significantly. The declining accuracy may be understandable in the case of greater number of clusters, but this also may be influenced by the distribution of the variables. Hence, general inferences cannot be deduced from these results.

An additional question is to measure the effect of the single variables which has a significant effect. How large is this effect? How much the change – which can be experienced in the data – is explained by the variables? There are more possibilities like $\eta 2$, adjusted $\eta 2$ and $\omega 2$ statistics (among others), which are able to solve this question. They compare the square of differences explained by the single factor with the all or rather the not explained square of differences [Vacha-Haase, Thompson 2004]. Among these statistics the $\eta 2$ can be interpreted in a percentile form. All the three statistics show that the model›s type (MOD) has the largest effect to the accuracy (Tab. 3), while the second most important is the number of the variables (NOV). Taking into consideration the interaction between two variables there is no as a great effect like in the case of MOD and NOV (e.g. $\eta 2$: $0.1 < 0.29 < 0.33$). This difference can be seen mostly in the case of the $\omega 2$ indicator (Tab. 3).

Conclusions

The aim of the examination has been to call the attention for the usage of procedures in the field of clustering, which can handle ordinal variables without distance measurement (e.g. Euclidean distance) and leads to a significantly more accurate result furthermore. This procedure is the Latent Class Analysis, with which one can get the groups not through defining the distance between the individuals, but by the help of principle of maximum likelihood.

It is done with the estimation of population's parameters, which is a procedure with large number of computations. However, the calculation time of models was between 1-2 minutes in case of an average personal computer. To tell the truth the Twostep algorithm was much faster (in comparison with poLCA), however it was developed directly for processing databases implying a lot of observation units.

Bibliography

Andrews Rick L., Brusco M., Currim I.S., Davis B. 2010: An Empirical Comparison of Methods for Clustering Problems: Are There Benefits from Having a Statistical Model? *Review of Marketing Science,* vol. 8, Article 3.

Bacher J., Wenzig K., Vogler M. 2004: SPSS TwoStep Cluster. A First Evaluation. Technical report. Universität Erlangen Nürnberg, Lehrstuhl für Soziologie, Nürnberg.

Chiu T., Fang D., Chen J., Wang Y., Jeris C. 2001: A robust and scalable clustering algorithm for mixed type attributes in large database environment. [In:] Proceedings of the 7th ACM SIGKDD international conference in knowledge discovery and data mining. Association for Computing Machinery, San Francisco, CA, 263-268.

Goodman L. 1974: The analysis of systems of qualitative variables. American Journal of Sociology. Vol. 79, 1179-1259.

Hubert L., Arabie P. 1985: Comparing partitions. *Journal of Classification*, vol. 2(1), 193-218.The SPSS TwoStep Cluster Component: A Scalable Component Enabling More Efficient Customer Segmentation. 2001: SPSS Inc. Technical report, Chicago, IL, [http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf], download 2012.01.13.

Linzer D. A., Lewis J. 2007: poLCA: Polytomous Variable Latent Class Analysis. R package version 1.1, [http://userwww.service.emory.edu/~dlinzer/poLCA], download 2012.01.13.

R Development Core Team. 2011: R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, [http://www.R-project.org], downloaded 2012.01.13.

Rand W.M. 1971: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* vol. 66, 846-850.

Vacha-Haase T., Thompson B. 2004: How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology,* vol. 51(4), 473-481.

Wedel M., Kamakura W.A. 2000: Market segmentation: Conceptual and methodological foundations (2nd ed.), Boston, MA: Kluwer Academic Publishers.

Zhang T., Ramakrishnon R., Livny M. 1996: BIRCH: An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD Conference on Management of Data. 103-114, Montreal, Canada.

Streszczenie

Przeprowadzono analizę i porównanie metod grupowania wykorzystujących modele oraz tych niekorzystających z metod modelowych. Stwierdzono, iż metody nieuwzględniające miar odległości (np. odległości Euklidesowej) pozwalają na osiąganie dokładniejszych wyników.

Correspondence address:

Ferenc Ruff, PhD, assistant lecturer
Szent István University
2100 Gödöllő, Hungary
Páter K. u. 1.
e-mail: ruff.ferenc@gtk.szie.hu