QUANTITATIVE METHODS IN ECONOMICS
Vol. XVI, No. 2, 2015, pp. 7 – 12

DETERMINING THE NUMBER OF CLUSTERS FOR MARKETING BINARY DATA

Jerzy Korzeniewski
Department of Statistical Methods, University of Lodz
e-mail: jurkor@wp.pl

Abstract: In the article a new way of determining the number of clusters was proposed focused on data made up of binary variables. An important application aspect is that the data sets on which the new formula was investigated were generated in the way characteristic for the marketing data following the work of Dimitriadou et al. [2002]. The new formula is a modification of the Ratkowsky-Lance index and proved to be better in some respects than this index, which was the best in the mentioned research. The modification proposed is based on measuring the quality of grouping into the predicted number of clusters and running the same index on the twice smaller set of objects comprising dense regions of the original data set.

Keywords: cluster analysis, binary data, number of clusters index, market segmentation

INTRODUCTION

Predicting of the number of clusters

One of very important parts of cluster analysis (unsupervised learning) is to find out how many clusters there should be in a data set. Obviously, this task is closely related to other cluster analysis tasks e.g. selection of variables and grouping of objects, however, the subject of selecting the proper number of clusters has attracted much interest which resulted in dozens of different proposals of indices or stopping rules. Milligan [1985] was probably the first to carry out a thorough investigation of more than two dozens of different indices but the research was concentrated on continuous variables data sets and it took place 30 years ago. Since that time many new proposals were published and the task has been directed to different targets related to e.g. different variable measuring scales. As far the binary variables are concerned a good examination was carried out by Dimitridou et al.

[2002]. The conclusion from this research is in favour of the Ratkowsky-Lance index which turned out to be better than other indices. Therefore, in order to carry out a new research on similar data sets this index was applied as the reference point. From a couple of newer proposals, the Fang and Wang index [2012] was also used in this article.

Binary marketing data

Binary marketing data specificity consists in a number of variables being correlated (or not) to create separate groups of variables. The whole data set consists of a couple of groups of such variables. In this research we followed the scheme suggested by Dimitraidou et al. [2002] in which every data set is described by twelve binary variables composed into four groups of different or equal numbers of variables. An example of such data pattern is presented in Table 1.

Table 1. An example of binary marketing data pattern, twelve variables in four groups

	Group1			Group2			Group3			Group4		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
Cluster1	H	H	H	H	H	H	L	L	L	L	L	L
Cluster2	L	L	L	L	L	L	H	H	H	H	H	H
Cluster3	L	L	L	H	H	H	H	H	H	L	L	L
Cluster4	H	H	H	L	L	L	L	L	L	H	H	H
Cluster5	L	L	L	H	H	H	L	L	L	H	H	H
Cluster6	H	H	H	L	L	L	H	H	H	L	L	L

Source: Dimitriadou et al. [2002]

The idea of this example is to present connections between groups of respondents and groups of questions in a questionnaire. The symbol H stands for the high probability of value 1 on a given variable and the symbol L stands for the low probability of 1. Obviously, the number of variables in each group, their correlation within the group, the level of H and L will be varied (see experiment description for details).

INDICES OF THE NUMBER OF CLUSTERS

Out of the multitude of the number of clusters indices which one can find in the literature we picked up as the reference point one that came the best in the Dimitriadou research i.e. the Ratkowsky-Lance index given by the formula

$$RL = \frac{mean\left(B/T\right)}{\sqrt{k}}, \tag{1}$$

where B stands for the sum of squares between the clusters for each variable, T stands for the total sum of squares for each variable and k stands for the number of groups into which the data has to be previously grouped by means of some grouping method.

The mean in the numerator of formula (1) is taken across all single variables. The value of *k* maximizing *RL* should be selected as the number of clusters prediction.

In order to include in the research some newer proposals we chose the Fang-Wang [2012] index based on the bootstrap method. This index is defined in the following way. We draw independently *B* bootstrap samples

$$X_b, Y_b, \quad b = 1,\ldots,B, \tag{2}$$

With the symbol $\Psi_{Xb,K}$ we denote the grouping of sample X_b nto *k* clusters. Then we define the distance of two groupings/divisions with the formula

$$\begin{aligned} &d\left(\Psi_{Xb,K}, \Psi_{Yb,K}\right) = \\ &= \frac{1}{n^2} \sum_{i,j=1} \left| I\left(\Psi_{Xb,K}(x_i) = \Psi_{Xb,K}(x_i)\right) - I\left(\Psi_{Yb,K}(x_j) = \Psi_{Yb,K}(x_i)\right) \right|, \end{aligned} \tag{3}$$

where *I* stands for the function assuming value 1 if the condition in the brackets is met. This distance measure has easy and intuitive interpretation. The final step is to define a measure of instability of divisions given by the formula

$$s_B(\Psi, K, n) = \frac{1}{B} \sum_{b=1}^{B} d\left(\Psi_{Xb,K}, \Psi_{Yb,K}\right). \tag{4}$$

The value of *k* this time minimizing the right hand side of formula (4) should be selected as the number of clusters prediction. All parameters necessary for the above formulas will be specified in the experiment description in the fourth chapter.

Some interesting recent proposals were given by Tibshirani et al. [2010] but they seem to be dedicated rather for special cases with the number of features being much bigger than the number of objects.

NEW INDEX PROPOSAL

We will try to propose a new index of the number of clusters which consists in the modification of the Ratkowsky-Lance index. The modification will involve two independent steps. One will be devoted to limiting the use of the original Ratkowsky-Lance index to half of the objects of a given data set belonging to "dense regions" of the data set. The other step will consist in measuring the quality of a data set division into a predetermined number of clusters.

Finding "dense regions" is a common concept in cluster analysis. The idea behind it is that limiting ones research to these regions usually gives more pronounced results in comparison with that of the whole data set. A popular technique of defining such regions is a sequential procedure working in the following way. The first object picked up is the one which has the smallest distance to its 20th nearest neighbor. This object is removed from the data set, all pairwise distances are

computed again and the second object picked up is the one with the smallest distance to its 20th nearest neighbor. We continue this process until we pick up half of the data objects. Obviously, the number 20 may be changed, for smaller data sets it is usually 5, but for the kind of sets used in our experiment (about 5000 objects) number 20 seems the proper choice.

Measuring the quality of a data set division or grouping is another task which can be performed in a number of ways. In our experiment we will use the following approach. Let us define a measure of the quality of a data set division into two clusters (which we will call the primary division). We choose all objects belonging to the smaller cluster and half of objects belonging to the bigger cluster and we divide these objects once again into two clusters using the same grouping method. The measure of the quality of the primary division will be given by the value of the adjusted Rand index [e.g. Gatnar and Walesiak 2004] as a similarity measure of both divisions. In the number of clusters prediction process, the data set is divided into different numbers of clusters, therefore, to use our measure we will apply it to every pair of clusters into which the data set was divided. For example, if the data set was grouped into 5 clusters we will get 10 measures of the quality of separation of every pair of clusters. Ideally, the value of 1 of the measure is desirable i.e. such value confirms that the division was well done or that the two clusters being assessed are perfectly separable. Formally, if anyone of the 10 values is close to zero i.e. very small it proves that in the division there is at least one pair of clusters which is badly separated. However, it only takes place in the case of very clear cluster structures that all pairs of clusters have division quality close to 1. Therefore, as the final measure of the division of the data set into any number of clusters, we will use a simple arithmetic mean across all pairs of clusters.

The new index formula is a modification of the Ratkowsky-Lance index the idea of which is to apply this index twice. Firstly to the whole data set and, secondly, to half of the data set representing dense regions. Subsequently, if the two instances return different numbers of predicted clusters, we will choose one of them. As we have to decide between from 2 to 10 clusters (see experiment description) we will concentrate our attention on the initial number of clusters i.e. 2, 3, 4 and 5. When the quality of these divisions (of the whole data set) is good we will use the prediction based on the whole data set. If the quality of these initial divisions is bad we will use the prediction based on the denser half of the data set. The logic behind such approach is that when divisions into smaller number of clusters are of bad quality the Ratkowsky-Lance index has a tendency to overestimate the predicted number of clusters. To be precise and not to search for thresholds taken from out of blue, we will use the value of 0.5 as the limiting value deciding about the divisions below this value being judged as bad divisions. Thus, the whole modification can be stated in the form of the following algorithm.

- Divide the whole data set into 2, 3, …, 10 clusters.
- Find the denser half of the data set using the technique of the 20^{th} closest neighbor.
- Divide the denser half into 2, 3, …, 10 clusters.
- If the measure of the quality of the whole data set division into 4 clusters is above 0.5 take the prediction of the Ratkowsky-Lance index based on the whole data set.
- If the measure of the quality of the whole data set division into 4 clusters is below 0.5 take the prediction of the Ratkowsky-Lance index based on the denser half of the data set.

EXPERIMENT DESCRIPTION

In order to evaluate the new index we carried out the following experiment. We generated 162 data sets according to the pattern described in chapter 2. We used the *bindata* package available in R language. The data sets generated were diversified with respect to the following parameters.

- Probability; for H there are 3 variants: 0.9, 0.8, 0.7 and for each variant respectively, for L there are 3 variants: 0.1, 0.2 0.3.
- Correlation inside groups of variables; there are 3 variants: uncorrelated variables, variables correlated with moderate strength (0.4), variables correlated with big strength (0.8).
- Number of clusters; 3 variants: 4, 5, 6.
- Numbers of objects in the clusters: 3 variants: (1000, 1000, 1000, 1000, 1000, 1000), (2000, 500, 1000, 700, 700, 1100), (3000, 300, 1000, 500, 700, 500).
- Number of variables within groups; 2 variants: (3, 3, 3, 3), (5, 4, 2, 1).

We ran the Ratkowsky-Lance index, the Fang-Wang index and the new proposal index using the *k*-means grouping method. The *k*-means grouping was done for a random choice of starting points, repeated 50 times, from which the result with the smallest distance measure was chosen. For the Fang-Wang index we used *B* equal to 50. The number of possible clusters from which the algorithms were choosing ranged from 2 to 10 clusters. In order to assess the efficiency of each index, out of many possible criteria, we used the percentage of properly predicted clusters as well as the percentage of errors equal to 1 and the percentage of bigger errors. In the literature one can find a couple of other criteria e.g. proper cluster recovery or correct dominant recovery. However, if one uses a mish-mash criteria the results are sometimes blurred because some criteria return different results than other criteria and does not get any clear conclusions.

RESULS AND CONCLUSIONS

The Fang-Wang index performed poorly achieving about 25% of correct predictions, therefore we will limit our conclusions to the other two compared indices. The two indices agreed in 50% of cases. Other results are given in Table 2. The new index achieved better overall performance as far as correct predictions are concerned (44% to 32%) with almost equal percentage of going wrong by 1 cluster.

Table 2. Results for the Ratkowsky-Lance index and the new index

	Performance measure	Overall performance	Number of clusters			Probability			Number of group variables	
			4	5	6	0.9	0.8	0.7	(3,3,3,3)	(5,4,2,1)
Ratkowsky-Lance	Correct hits	.32	.43	.20	.35	.41	.44	.12	.49	.16
	Error = 1	.27	.15	.43	.23	.37	.24	.19	.34	.19
New index	Correct hits	.44	.52	.56	.25	.35	.50	.48	.44	.44
	Error = 1	.28	.24	.26	.38	.33	.33	.21	.31	.28

Source: own research

The new index was also better in most subcategories apart from the sets with 6 clusters (the new index lost 25% to 35%), clear cluster structures (the new index lost 35% to 41%) and apart from the group with uniform numbers of variables (the new index lost 44% to 49%). The basic reason for the poorer performance of the Ratkowsky-Lance original index seems to be its poor results (only 16% ! of correct hits) for the data sets in which some groups of variables have much smaller numbers of variables than other groups as well as very poor result (only 12% ! of correct hits) for blurred cluster structures. In conclusion we can state that the new proposal is more robust to unwelcome conditions.

REFERENCES

Dimitriadou E., Dolničar S., Weingessel A. (2002) An examination of indexes for determining the number of clusters in binary data sets, Psychometrika, Vol. 67, Issue 1, pp. 137-159.

Fang Y., Wang J. (2012) Selection of the number of clusters via the bootstrap method, Comput. Statist. Data Anal. 56, pp. 468–477.

Gatnar E., Walesiak M. (2004) Metody Statystycznej Analizy Wielowymiarowej w Badaniach Marketingowych, Wydawnictwo AE we Wrocławiu, pp. 334-36.

Leisch F., Weingessel A., Hornik K. (2015) *bindata* package manual.

Milligan G. W., Cooper M. C. (1985) An examination of procedures for determining the number of clusters in a data set, Psychometrika 50, 159–179, pp. 137-159.

Tibshirani R., Witten D. (2010) A framework for feature selection in clustering, Journal of American Statistical Association, 105(490), pp. 713–726.