STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 1–10

NEW METHOD OF VARIABLE SELECTION FOR BINARY DATA CLUSTER ANALYSIS

Jerzy Korzeniewski[1]

ABSTRACT

Cluster analysis of binary data is a relatively poorly developed task in comparison with cluster analysis for data measured on stronger scales. For example, at the stage of variable selection one can use many methods arranged for arbitrary measurement scales but the results are usually of poor quality. In practice, the only methods dedicated for variable selection for binary data are the ones proposed by Brusco (2004), Dash et al. (2000) and Talavera (2000). In this paper the efficiency of these methods will be discussed with reference to the marketing type data of Dimitriadou et al. (2002). Moreover, the primary objective is a new proposal of variable selection method based on connecting the filtering of the input set of all variables with grouping of sets of variables similar with respect to similar groupings of objects. The new method is an attempt to link good features of two entirely different approaches to variable selection in cluster analysis, i.e. *filtering* methods and *wrapper* methods. The new method of variable selection returns best results when the classical *k*-means method of objects grouping is slightly modified.

Key words: cluster analysis, market segmentation, selection of variables, binary data, k-means grouping.

1. Introduction

Feature selection is probably the most important stage of cluster analysis just like in many other parts of statistics. The results of variable selection determine significantly the final results of cluster identification and if received incorrectly may render it impossible to identify any clusters. The task of variable selection in cluster analysis has probably been highlighted by the well-known article by Carmone et al. (1999) in which the HINoV method was proposed. Although in this article the authors mention some earlier attempts to approach the task of variable selection, they assess them as absolutely infeasible in application to empirical data. After 1999 several methods or algorithms for variable selection were proposed, however, most of them were meant rather for strong scales on

[1] University of Lodz, Poland. E-mail: jurkor@wp.pl.

which the variables are measured. A good evaluation of some of them is given in Steinley et al. (2008) and in Korzeniewski (2012). Some of these methods allow even for a form of statistical inference like, e.g. Raftery and Dean (2006) method. As far as weaker measurement scales are concerned, e.g. binary data sets, it is not easy to find any well-performing methods. The methods developed by Brusco (2004), Dash et al. (2000) and Talavera (2000) should be mentioned as the ones to be investigated and assessed. A particular problem of cluster analysis of binary data arises when one is confronted with the task of market segmentation. A characteristic feature of this type of data is the existence of a couple of groups of binary variables with possible pairwise correlation within the groups. Such data usually are confronted with when carrying out statistical research of large numbers of clients on the market. This kind of research is very often performed in the form of a questionnaire comprising several questions with possible binary answers. Dimitriadou et al. (2002) proposed a way of simulating this kind of binary data for the task of determining the number of clusters. Their *bindata* package as recent as 2015 is freely available in R language. One very characteristic feature of the marketing type of binary data is its relatively big size – the number of possible objects is at least several thousands. The conclusion, therefore, is that one rather has to use partitioning methods of objects grouping, one cannot use, e.g. agglomerative methods.

The objective of this article is to propose a new and efficient method for the earlier stage of cluster analysis, i.e. variable selection on the marketing type of binary data and to assess the efficiency of this method in comparison with other existing methods. The article is organized as follows. In the next part an overview of the methodology of three methods is presented with possible hints of their applicability to the task in the context of marketing data. Part three presents a proposal of the new method. Part four includes an empirical evaluation of the new method and other methods. The final fifth part contains conclusions and prospects for future research.

2. Overview of variable selection methods

The number of variable selection methods in cluster analysis is quite large comprising several proposals, however, methods which were constructed for strong measurement scales (predominantly continuous variables) do not perform well for binary variables or cannot be applied at all. This phenomenon is quite common across all statistical methods. Therefore, we limit our examinations to the three methods described in this chapter which were constructed for nominal scales, some of them especially for a binary scale.

The Brusco method of variable selection dedicated for binary variable consists of the following steps:

1) For arbitrary subset of the set of all variables group all data set objects 500 times in the predetermined number of clusters and remember the sum of average distances inside the clusters.

2) Add a single new variable to this subset if the new sum (with the new variable) of average distances is smaller than the previous one (the very subset), i.e. $Z_{\min} < Z_{B2}$.
3) Stop the process of variable adding if

$$Z_{B2} > Z_{\min} + \delta \frac{M}{4} \tag{1}$$

where M is the number of data set numbers and $\delta \in (0,1)$ is a parameter to be fixed intuitively.

There are some major doubts which can be raised about this method. Firstly, Brusco uses the classical form of *k*-means grouping stating that it renders good results. The results depend on the type of data used in the experiment. If the data were slightly more obscure (clusters less distinct) the results could be much worse because the classical *k*-means is not well suited for binary variables as in the first loop there are many draws (equal Sokal-Michener distances) and it is impossible to say into which direction a given object should "go". Secondly, the number of clusters into which the sets have to be partitioned is the proper one and Brusco advocates that there are "excellent" ways of determining the number of clusters, mentioning and recommending the Ratkowsky-Lance index. The Ratkowsky-Lance index was investigated by Dimitriadou et al. (2002) on the marketing binary data (as well as by other authors) and the results are quite clear – it finds the proper number of clusters in about 70% of cases, sometimes going wrong by more than two clusters. If the number of clusters was established erroneously then the results would probably be worse. Thirdly, there is a question of the level of cluster separation which is discussed below.

Table 1. The pattern of Brusco binary data.

Number of clusters	Number of variables		
	4	6	8
4	1 0 0 1 1 1 1 0 0 0 1 1 0 1 0 1	1 0 0 1 1 0 1 1 1 0 0 0 0 0 1 1 0 0 0 1 0 1 0 1	1 0 0 1 1 0 1 0 1 1 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 1 0
6	1 0 0 1 1 1 1 1 1 0 1 0 0 1 0 1 0 0 0 1 0 1 1 0	1 0 0 0 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 0 0 1 0 1 1 1 1 0 0 0 0 1 1 0	1 0 0 0 1 1 0 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 0 0 1
8	1 0 1 1 1 0 0 0 1 1 1 0 1 1 0 1 0 1 0 1 0 1 0 0 0 0 1 1 0 0 0 1	1 0 0 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 0 1 0 0 1 0 0 1 1 0 0 1 0 0 1 1 1 0 0 0 1 0 0 1	1 0 0 1 1 1 0 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0 1 0 1

Source: Brusco (2004).

The outlined above method was evaluated by Brusco on the data sets the skeleton of which is given in Table 1. Obviously, the data was varied with respect to the size of clusters, the number of objects in data sets, the level of cluster separation, etc. A natural question arises: what is the difference between data sets of this type and the ones generated by Dimitriadou et al. (see section 4)? The answer seems to be that the major difference lies in the level of cluster separation. Brusco allows only for 4% (at the worst case) of 1 being changed into 0 or vice versa. The way of defining cluster separation is entirely different in the work of Dimitriadou et al. (2002). It seems, however, that their levels of, e.g. 0.8 for 1 and 0.8 for 0, allow for much less clear cluster structure, to say nothing of the levels of 0.7 and 0.3, respectively.

The method proposed by Talavera consists in using the formula

$$Kor(v_M) = \frac{\sum_v \sum_j P(x_v = a_{vj}) \sum_{j_M} \left(P^2(x_{v_M} = a_{vj_M} \,|\, x_v = a_{vj}) - P^2(x_{v_M} = a_{vj_M}) \right)}{\left| \{ v \,|\, v \neq v_M \} \right|} \tag{2}$$

for arranging all variables in descending order with respect to the strength of correlation between variable v_M and the remaining variables. In formula (2) the symbol a_{vj} stands for the *j*-th variant of *v*-th variable and the formula was derived with the use of the Bayes theorem starting from the maximization of a measure of the quality of the division of the data set into a predetermined number of clusters. It seems that we can assess this method at this stage basing our judgement on the evaluations which can be found in the literature. In order to apply the Talavera method to a particular data set one has to use the COBWEB algorithm (or similar based on a hierarchical tree). All applications to be found (e.g. Devaney et al. (1997)) analyse small data sets of not more than a couple of hundred objects (e.g. heart disease UCI data set and LED UCI data set). It is not feasible to apply this method to the whole data sets of the marketing type (one rather has to use partitioning methods) unless one tries to draw small samples and somehow unify the results. Besides, the number of clusters has to be known.

The Dash and Liu method is a very general method which can be applied to any measurement scale because it is based on the analysis of the data set entropy. The smaller the entropy (for different combinations of variables used) the better it is for the considered set of the variables with respect to the strength of evidence on possible cluster structure. The entropy of the data set is measured with the formula

$$E = -\sum_{x_1, x_2} \left[S(x_1, x_2) \log S(x_1, x_2) + (1 - S(x_1, x_2)) \log(1 - S(x_1, x_2)) \right] \tag{3}$$

where $S(x_1, x_2)$ stands for the similarity of two objects being a simple transformation of the distance between these two objects. It seems that we can assess the applicability of this method at this stage taking into consideration its basic characteristics. Firstly, the entropy-based method only allows for the ordering of all variables with respect to their importance to a possible cluster structure. If this ordering is done incorrectly there are no chances of the correct selection of variables. This seems to be a major drawback. Secondly, one needs some kind of criterion as to where to divide the sequence of ordered variables. The authors suggest a criterion based on the results of objects grouping. However, as in the case the Brusco method, the number of clusters has to be predetermined. It is possible to assess the entropy of the data set on any chosen subset of the set of all variables and pick up the best one. However, this leads to the necessity of examining all possible subsets of variables.

3. New method formulation

To make the presentation of our proposal as clear as possible let us start from dividing this method into two stages:

- Stage 1. Filtering stage which consists in grouping all variables into classes of similar variables with respect to some kind of correlation measure.
- Stage 2. Wrapper stage which consists in possible grouping of the classes of variables received in stage 1 with respect to the similarities of grouping of the data set objects.

Any method consisting of the two steps given above is not going to work properly if one uses classical techniques like, e. g. coefficient of linear correlation in stage 1 or

k-means clustering in stage 2 due to well known drawbacks of these methods when applied to binary data. However, if we use more versatile measures the method is going to work very well.

Thus, in stage 1 we will use the *distance based correlation* (Korzeniewski, 2012) between two sets *A*, *B* of variables given by the formula:

$$DBC(A,B,l) = \frac{\frac{1}{l}\sum_{i=1}^{l} d_i^A d_i^B - \bar{d}^A \bar{d}^B}{s^A s^B}, \tag{4}$$

where $1 \le l \le n$ denotes the number of observation pairs drawn without replacement from all pairs of observations; d_i^A, d_i^B denote distances for *i*-th pair of objects based on the variables from sets *A*, *B*, respectively; $\bar{d}^A$, $\bar{d}^B$, s^A, s^B denote arithmetic means and standard deviations computed from all *l* distances on both sets of variables, respectively. This kind of correlation measure is extremely useful when applied in cluster analysis (Korzeniewski,

2012) because if there is a cluster structure and both sets of variables *A* and *B* participate in creating it, then any substantial changes in distances between objects on set *A* should cause changes of distances on set *B*. To fix all technicalities let us establish that we will apply formula (4) only to sets *A* and *B* consisting of single variables and *l*=20 with the value of $DBC(u,v)$ (*l* is skipped) being the arithmetic mean from 100 repetitions.

In stage 2 we have to use some kind of partitioning algorithm because the data sets are too big for agglomerative algorithms. The most popular and applied in virtually any comparative simulation study is the *k*-means clustering. However, in the case of binary data sets, it cannot be used under no pretence whatever. Firstly, as in the case of the methods described in section 2, we would have to specify the number of clusters *k* which would give no advantage over the other methods. Secondly, one of the basic drawbacks of *k*-means partitioning of binary data is its ambiguity caused by huge percentage of draws in the first loop of the *k*-means partitioning. We propose the following partitioning process based on multiple *k*-means partitioning for *k*=2. We partition the whole data set into two clusters, then each of the two clusters is partitioned into two clusters and so on. Such a way of partitioning gives much better results being a cure for almost all *k*-means ailments. We only have to specify a stopping rule. It can be, e.g. the minimum cluster size, however this would be a new parameter, nowhere to be found. A better way is to specify a threshold of a reasonable quality of partitioning into 2 clusters. There are many measures of data division quality, e.g. the ones based on replication techniques. Another stopping rule can refer simply to the depth of the partitioning process. In the case of the binary marketing data this way is absolutely sufficient since due to multiple partitioning of the same (or very similar) data sets being subsets of the whole data set and the random character of the initialization of *k*-means partitioning, as well as a small number of clusters into which we want to segment the market, it is enough if we establish the depth to be 3. To be more precise, we partition the whole data set into 2 clusters, each of which is partitioned into two clusters, each of which is likewise partitioned. There is one another justification for this relatively small depth of partitioning, namely the value of the threshold from which we will decide that two divisions made on two different sets of variables are similar. The measure of the quality of division will be the adjusted Rand index (Hubert, 1985) which usually assumes values from interval (0,1). We set the threshold value to be 0.15. This is a small number as far as demanding high similarity of divisions is considered. However, it is sufficient, as two different binary variables with equal and random distributions of their variants never returned at least one value of the Rand index greater than 0.15 in 1000 random simulations (*k*-means, *k*=2, random starting points). Therefore, the threshold of 0.15 seems to be a very mild one and at the same time a rigorous one. As the threshold is very small we do not have to seek very intently for two very similar partitions, that is why we can stop our multiple partitioning at the depth of 3.

Summing up the above considerations we propose the following steps:

1. Group all variables into classes of variables such that in each class, for every variable *v* there is variable *u* such that $DBC(u,v) > 0.1$.
2. Merge two different groups of variables resulting from step 1 if the value of the adjusted Rand index between any of the 8 divisions of the data set made on one group of variables and any of the 8 divisions made on the other group of variables exceeds 0.15.
3. Repeat step 2 until no merges can be made.
4. Consider all single variables to be noisy variables, i.e. not participating in creating cluster structure and discard them.
5. Consider each class of variables consisting of more than one variable to be important for cluster structure. If there is more than one such class, it suggests the existence of multiple cluster structures.

4. Simulation experiment

In order to assess the efficiency of the new method on binary marketing data 162 data sets were generated. We followed the pattern 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 2. The idea of this example is to present connections between groups of

Table 2. 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].

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 vary. We used the very recent *bindata* (Leisch et al., 2015) 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 great 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).

If there are less than 6 clusters we take into account only the initial clusters, i.e. the ones from the top of Table 2. All combinations of variants result in 162 data sets. In order to assess the efficiency of the new method in proper variable selection a similar number of noisy variables were added to each of 162 data sets. The noisy variables resulted in adding 8 uniformly distributed sets of observations coming from pairwise independent binary variables (equal probabilities for 1 and 0).

Every *k*-means partitioning in the new algorithm described in section 2 follows the classical form of this method, i.e. the two starting objects are randomly chosen, the procedure is repeated 100 times and we pick up the variant with the smallest sum of squared distances. The distance measure used is the Sokal-Michener distance.

5. Results and conclusions

The new method performed very well because it was almost perfect in 89% of data sets and absolutely wrong (returning 20 separate variables, i.e. discovering no cluster structure) in 11% of data sets. What is more, the wrong decisions comprised all 18 data sets (and none else set) with no correlation in the groups of variables creating cluster structure and the weakest variant of cluster separation, i.e. probabilities of 0.7 and 0.3 for high (H) and low (L) probability of 1, respectively. In other words, if there is at least a small hint of cluster structure existence (i.e. correlation between variables or decent levels of cluster separation), the new method is very likely to detect it. Other numerical characteristics of the results are as follows. In two cases (1.2% of data sets) the new method incorporated more than 4 noisy variables into the set of variables true for cluster

structure. In 9 cases (5.6% of data sets) the new method found 2 separate cluster structures, usually one "major" created by 8 or 9 variables and one "minor" created by 2 or 3 variables. It seems, however, that in these cases one could pursue the detecting similarities of divisions of data set objects in some other way than the techniques used, because one has more options to detect such similarities when there are a few variables in each of the two sets of variables. In 21 cases (13% of data sets), with the vast majority from the second type of data sets, i.e. the second (5, 4, 2, 1) case of the numbers of variables in each group, the new method missed one single variable (properly detecting 11 others). The new method allows for perfect variable selection (selecting 12 true variables and discarding 8 single remaining variables) in 65% of data sets.

REFERENCES

CARMONE, F., KARA, A., MAXWELL, S., (1999). HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables, Journal of Marketing Research, Vol. 36, No. 4, 501–510.

DASH, M., LIU, H., (2000). Feature selection for clustering, Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD), 110–121.

DEVANEY, M., RAM, A., (1997). Efficient feature selection in conceptual clustering, Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, 92–97.

DIMITRIADOU, E., DOLNICAR, S., WEINGESSEL, A., (2002). An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets, Psychometrika 67(1), 137–160.

HUBERT, L., ARABIE, P., (1985). Comparing Partitions, Journal of Classification 2.

LEISCH, F., WEINGESSEL, A., HORNIK, K., (2015). Bindata package manual.

KORZENIEWSKI, J., (2012). Selekcja zmiennych w analizie skupień. [The selection of variables in cluster analysis]. Nowe procedury, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

RAFTERY, E., DEAN, N., (2006). Variable selection for model-based clustering, Journal of the American Statistical Association , 101(473): 168–178.

STEINLEY, D., BRUSCO, M., (2007). Initializing k-means batch clustering: A critical evaluation of several techniques, Journal of Classification 24, 99–121.

STEINLEY, D., BRUSCO, M., (2008). Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures, Psychometrika 73, 125–144.

TALAVERA, L., (2000). Dependency-Based Feature Selection for Clustering Symbolic Data, Intelligent Data Analysis 4, 19–28.