A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Jerzy Korzeniewski[*]

MODIFICATION OF HINOV METHOD OF VARIABLE SELECTION FOR MULTIPLE CLUSTER STRUCTURE ANALYSIS

Abstract. The original HINoV method (*Carmone et al.,* 1999) is not robust to the presence of correlated unimodal and uniform variables among noisy variables (*e.g. Korzeniewski,* 2012). Moreover, HINoV can be applied only to a single cluster structure analysis. In the article, a modification is proposed consisting in grouping all variables (separately for each reference variable) into two classes. One of the classes consists of variables similar to the reference variable, the other consists of variables which are "less similar". Similarity between two variables is based on the similarity of the data set division into an established number of clusters (from 2 to 10) measured with the modified Rand index. We arrive at a zero-one matrix describing relations between every pair of variables. Then, a set of variables creating the same (the strongest) cluster structure is selected by means of a criterion optimizing the matrix division into four blocks. After completing the first stage selection one can search another cluster structure applying the same procedure to the set of remaining variables. The modification is assessed in a broad experiment based on 2250 data sets generated from the mixtures of normal distribution.

Key words: cluster analysis, variable choice, multiple cluster structures.

I. INTRODUCTION

It is widely acknowledged that not all variables characterising data set observations contribute the same weight to the data set cluster structure. Some are more important than others (true variables), some are less important and some may be an obstacle (masking variables) in detecting the data set cluster structure. In recent years quite a number of methods designed with the aim of choosing the best subset of variables describing the data set cluster structure was proposed. The proposal of HINoV (*Carmone et al.,* 1999) is by some researches considered a turning point of the task of feature selection in cluster analysis. The method presented in the original article is not a strictly statistical method because it is based on visual assessment of the scree plot, but in this publication the authors criticized all formerly developed methods. Moreover, HINoV, although very imprecise, gained wide recognition among statistical community and even some modifications were proposed e.g. *Brusco* and *Cradit*, (2001),

[*] Ph.D., Department of Statistical Methods, University of Lodz.

Steinley and *Brusco* M. (2007). No one, however, has made an effort to modify HINoV with the aim to analyse multiple cluster structures. In general, there are practically no methods of variable selection in the case of multiple cluster structures. Probably, the only work which addresses this problem was written by *Friedman* and *Meulman* (2004), but so far, their method was not implemented for practical use. The idea of the method presented in this article consists in replacing the TOPRI indicator used in HINoV by similar indicator computed separately for each variable. In consequence, in the next stage, we have to divide the matrix of variables similarities into blocks and pick up some variables which will represent the first set of "similar" variables. We repeat these stages until all variables are grouped into "similar" subsets. In the final stage we have to make a decision for each of the subjects on whether to qualify it as a subset creating cluster structure or not. A criterion used for that purpose is proposed. The details of the method are presented in the second chapter with the analysis of the method efficiency in the following chapters.

II HINOV MODIFICATION

While working out new methods of variable selection in the case of possible multiple cluster structures we have to pay attention to the following four main targets:

- possible existence of many cluster structures (main target);
- robustness to the existence of correlated unimodal or uniform variables;
- robustness to the number of noisy variables being large in comparison with the number of variables creating cluster structures
- robustness to the unknown number of clusters.

HINoV works in the following way. For each variable v we group data using e.g. k-means method (k has to be specified). Then we compute the $R(u,v)$ Rand index which is a measure of similarity of the two variables for which it is computed. Subsequently, we rank all variables according to the *TOPRI* indicator defined as:

$$TOPRI(u) = \sum_{u \neq v} RI(u,v) \tag{1}$$

The variables with highest values of indicator (1) are interpreted as the ones having strongest connection with the existing cluster structure. Such approach has a number of drawbacks (compare *Korzeniewski,* 2012) and obviously cannot be used to detect multiple cluster structures.

In view of the second and third target mentioned above we propose to get rid of the summation in formula (1), because, if noisy variables are numerous and correlated then it always results in high values of *TOPRI* for variables which may not create any cluster structure. Instead we will assess the similarity $RI(u,v)$ of two variables with the biggest of the Rand index values for the range of possible numbers of clusters from 2 to 10. Thus

$$RI(u,v) = \max_{k=2,..,10} \{RI(k,u,v)\}. \tag{2}$$

In the next step we arrange all $d \cdot d$ Rand indices in the form of a matrix

$$R = [R_{ij}] \tag{3}$$

where R_{ij} is the corrected Rand index between variables *i* and *j*. In order to group variables into subsets of more than two similar variables we will apply a sequential procedure. One step of this procedure consists in selecting subset *A* of variables which maximizes criterion

$$\sum_{i,j \in A} RI_{ij} - \sum_{\substack{i \in A \wedge j \notin \mathrm{A} \\ \vee\, i \notin A \wedge j \in \mathrm{A}}} RI_{ij} \tag{4}$$

Sequential application requires elimination of variables belonging to the subsets selected in steps prior to current step. After we have grouped all variables into similar subsets we have to make a decision for each subject whether to consider it as creating cluster structure or not. We propose a simple criterion which turned out to be quite effective in former studies (*Korzeniewski,* 2012). The criterion may be easily formulated verbally as the Rand index of the consistency of the data set division into two clusters with the division into two clusters of the data set subset consisting of the smaller cluster and 1/3 of the bigger cluster. We will denote the criterion with

$$R(A). \tag{5}$$

The idea of criterion (5) is presented in Figure 1 and Figure 2. If there is a cluster structure (Fig. 1) the first division into two clusters (the bigger represented with dark bright squares and the smaller represented with circles) should be consistent with the division into two clusters of the subset consisting of the smaller cluster (circles) and the closest 1/3 (dark squares) of the bigger

cluster. If there is no cluster structure (Fig. 2) there is no reason for the second division to break up the subset consisting of circles and dark squares in the way similar to the first division. In order to find out the value of criterion (5) one has to decide on the grouping method and the "depth" to which the divisions will be carried out. The depth bigger than 1 is necessary because if we try to divide the data set only once we will not get a big value of criterion (5) for the cluster structures of 6 or 7 (or more) clusters – the division of many clusters into two subset does not have to return clear "division gap". We used depth equal to 2 which has simple interpretation i.e. we calculate criterion (5) for the (first stage) division of the whole data set and (more deeply) we calculated criterion (5) for the division of each of the two clusters resulting from the first stage division.

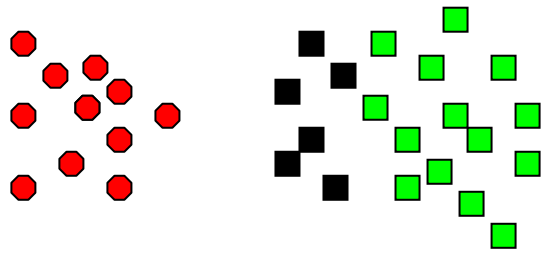

Fig. 1. Graphical representation of criterion (5) in the case of cluster structure

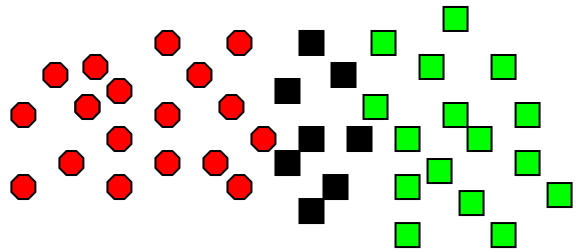

Fig. 2. Graphical representation of criterion (5) in the case without cluster structure

Taking into account all improvements mentioned above we may try the following modification:

Step 1. Group all data set observations independently for every variable into 2, 3, …, 10 clusters.

Step 2. Find matrix R according to formula (3).

Step 3. Find subset A of variables which maximizes criterion (4) from all possible subsets of the set of all variables.

Step 4. Find the value of $R(A)$ for depth of division equal to 2, i.e. the biggest of the three Rand indices measuring the data set first stage division into two clusters and the likewise divisions into two clusters of each of the two clusters from the first stage division.

Step 5. If there are some variables not included in any of the sets A found so far, go to step 3.

Step 6. The sets A which have $R(A) > 0.4$ qualify as sets creating cluster structure.

In the efficiency investigation from the following parts, in step 1, we will use standard k-means grouping with random choice of starting points, repeated 100 times with the minimal between clusters variance criterion to choose the final grouping. The distance measure will be the Euclidean distance. In step 2 we will use the corrected Rand index as described e.g. in *Gatnar, Walesiak* (2004).

III EFFICIENCY INVESTIGATION

In order to assess the efficiency of our modification we generated cluster structures according to the Steinley and Henson's OCLUS algorithm (2005) based on generating each cluster from the standard normal distribution (on each dimension). We generated cluster structures consisting of 200 data items, differing with respect to the following factors.

The first factor, the number of clusters in the data set was examined at five levels – 2, 3, 4, 6 and 8 clusters.

The second factor, number of items in clusters was examined at three levels: (a) an equal number of objects in each cluster; (b) 10% of objects and (c) 60% of objects in one cluster and the remaining objects equally divided across the remaining clusters.

The third factor, the number of true variables was tested at three levels – 2, 4 and 6.

The fourth factor, the probability of overlap between clusters on each true variable was tested at five levels – 0, 0.1, 0.2, 0.3, 0.4. The overlap was of the "chain" type (see *Steinley* and *Henson*, 2005) and so, on each dimension, there were k-1 pairs of overlapping clusters (k – number of clusters).

The fifth factor, the degree of within-cluster correlation had two variants: (a) the covariance matrix for each cluster was the identity matrix ; (b) each cluster had the same covariance matrix with ones on the diagonal and the off-diagonal elements drawn from a continuous distribution on the interval [0.3; 0.8].

The number of combinations is equal to 450, repeated 5 times results in 2250 cluster structures.

There is some ambiguity in assessing the efficiency in the case of multiple cluster structures. We used the following approach. All cluster structures were

kept and to every structure another was drawn randomly from all 2250 structures. In this way we got fairly smooth distribution of cluster structure types. Then, to every two-fold (i.e. consisting of two structures) cluster structure four noisy variables were attached: two independent uniform distributions on interval [0; 20] and two normal distribution with zero mean and covariance matrix with ones on the diagonal and 0.5 or 0.75 (randomly chosen) off the diagonal. Thus, every cluster structure was masked with four variables.

Standard measures usually used comprise two criteria (compare *Steinley and Brusco,* 2008): recall - the number of relevant variables in the chosen subset of variables divided by the total number of relevant variables and precision – the number of relevant variables in the chosen subset of variables divided by the total number of variables selected. Recall and precision are computed for every data set and the arithmetic means of these two measures are computed for all data sets. However, it is not clear how to use these measures in the presence of several cluster structures. We applied the following approach. All selected subsets of variables were ordered decreasingly with respect to their strength (i.e. the $R(A)$ value), one-variable subsets were eliminated. From this sequence only the two first subsets of variables were considered (on condition that their strengths were greater than 0.4). For each of the two known cluster structures best variants of precision and recall (obtained from the two strongest cluster structures) were found. The arithmetic means of these measures were subsequently computed as the final measure. An alternative approach might be considering all cluster structures with strength exceeding 0.4 (as those are considered structures detected in our method's formulation), however, that would require additional efficiency measures apart from recall and precision.

IV RESULTS AND CONCLUSIONS

In Table 1 the final means of recall (upper numbers in each row) and precision (lower numbers) are presented. The numbers from the table and the overall performance allow to draw the following conclusions.

Table 1. Recall and precision of the new method with respect to cluster structure overlap

	Overlap 0	Overlap 0,1	Overlap 0,2	Overlap 0,3	Overlap 0,4
No correlation within clusters	1.0 1.0	0.63 0.71	0.57 0.66	0.42 0.53	0.43 0.55
Correlation within clusters	1.0 1.0	0.59 0.70	0.56 0.65	0.41 0.48	0.40 0.50

Source: own investigations.

1. The method proposed is strictly statistical i.e. it is based on numerical measure and not on e.g. visual examination.

2. The method results are good in the cases of distinct cluster structures (overlap equal to 0 or 0.1) and worse for weaker structures, however, the results cannot be compared with other researches as no investigation of multiple cluster structures is known to the author.

3. The method is not feasible for data sets described by a large number of variables since we have to check (step 3) all possible subsets of variables, although checking amounts only to computing sums of a couple of dozen real numbers which is a fast procedure. This drawback does not seem to be very vital as for cluster structures created by more than 10 variables and not so distinct as to be analysed on marginal histograms, the analysis of individual variables is of limited use anyway.

4. The method proposed may be modified with respect e.g. to the algorithm of selecting subsets of „similar" variables.

5. The method proposed may be modified for the need of the most general formulation of cluster analysis variable selection problem i.e. the case of multiple cluster structures created by sets of variables not necessarily disjoint.

REFERENCES

Brusco M. J., Cradit J.D., (2001), A variable-selection heuristics for K-means clustering, Psychometrika 66.

Carmone F. J. Jr., Kara Ali, Maxwell S. (1999), *HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables*, Journal of Marketing Research, Vol. 36, No. 4.

Friedman J., Meulman J. (2004), *Clustering Objects on Subsets of Attributes*, Journal of the Royal Statistical Society, Series B 66.

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

Korzeniewski J. (2012), *Metody selekcji zmiennych w analizie skupień. Nowe procedury,* Wydawnictwo Uniwersytetu Łódzkiego.

Steinley D., Brusco M. (2007), *A new variable weighting and selection procedure for K-means cluster analysis*, Psychometrika 66

Steinley D., Brusco M. (2008), *Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures*, Psychometrika 73 No. 1.

Steinley D., Henson R. (2005) OCLUS: *An analytic method for generating clusters with known overlap*. Journal of Classification, 22.

Jerzy Korzeniewski

MODYFIKACJA METODY HINOV SELEKCJI ZMIENNYCH W ANALIZIE WIELOKROTNYCH STRUKTUR SKUPIEŃ

Oryginalna metoda HINoV jest zupełnie nieodporna na występowanie wśród zmiennych zanieczyszczających strukturę skupień zmiennych skorelowanych jednomodalnych lub równomiernych. Ponadto HINoV można stosować tylko w przypadku jednej struktury skupień.

W referacie zaproponowana jest modyfikacja polegająca na tym, by, oddzielnie, dla każdej ustalonej zmiennej, grupować zmienne w dwie klasy zmiennych podobnych i niepodobnych do niej w sensie podobieństwa podziału zbioru danych na daną liczbę skupień (od 2 do 10). Otrzymujemy wówczas macierz zerojedynkową opisującą związki pomiędzy każdą parą zmiennych. Następnie, podzbiór zmiennych tworzących tę samą (najsilniejszą) strukturę skupień wybierany jest za pomocą kryterium optymalizującego podział macierzy na cztery bloki. Po wybraniu zmiennych tworzących jedną strukturę skupień można, w dalszym kroku, wybierać zmienne tworzące następną strukturę skupień spośród zmiennych, które nie zostały wybrane w pierwszym kroku. W celu selekcji właściwego bloku macierzy stosowane jest kryterium stabilności podziału zbioru danych oparte na wielokrotnym losowaniu połowy zbioru i porównywaniu podziałów otrzymanych przy pomocy metody *k*-średnich. Modyfikacja oceniona jest w obszernym eksperymencie symulacyjnym na 2250 zbiorach danych wygenerowanych w postaci mieszanin rozkładów normalnych.