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 235, 2010

Jerzy Korzeniewski[*]

INVESTIGATION OF THE EFFICIENCY OF A NOVEL ALGORITHM FOR THE CHOICE OF VARIABLES IN CLUSTER ANALYSIS ON REAL WORLD DATA SETS

Abstract. In the paper, we investigate the efficiency of an Author-proposed algorithm for the choice of variables in cluster analysis on real world data sets. The assessment of this algorithm on synthetic data sets in the form of the mixtures of normal distributions was the subject of other paper – the algorithm turned out to be well working. The idea of the algorithm is to pick up as true these variables whose variance does not get so small as the variance of masking variables in the one-step mean shift procedure of data set observations.

Key words: cluster analysis, variable choice.

I. METHODOLOGY

In order to present the idea of the new method let us consider an examplary data set. The set (left set in Fig. 1) consists of two clusters arranged in such a way that only the second varaiable should be considered as true – the first vraiable may be treated as masking because it has no impact on assigning points to clusters. The one step mean-shift procedure applied to the window size of the arithmetic mean euclidean distance between pairs of points works so that it moves each point x to the mean of all the points lying within the window of the size considered and centered at x. It is easy to see that that if we apply such a procedure to this set the result will be two very tight clusters centered closely to the original clusters centres (right set in Fig. 1). The variance of the first variable will be close to zero while for the second varaiable it will be significantly larger because, with respect to the second variable, the cluster centres (both original and after shift) are widely spaced. Thus, the idea of the method is very simple: *the smaller the variable's variance after shift the less likely it is to be a true varaible and the morelikely to be a masking one.*

[*] Doctor, Department of Statistical Methods, University of Łódź.

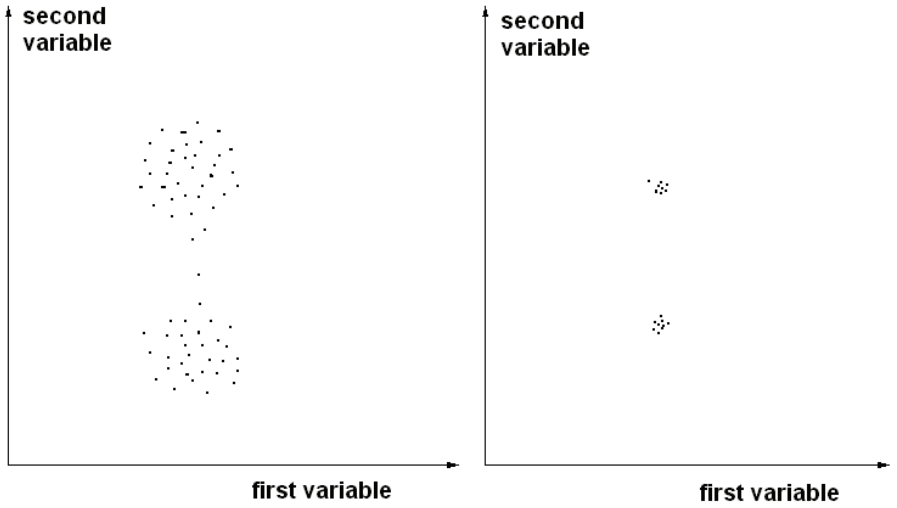


Fig.1 Examplary two dimensional set consisting of two clusters. The first variable is a masking variable.

The new method of choosing variables in cluster analysis works well for data sets being formed by the mixtures of normal distributions (see Korzeniewski 2009). In this paper we intend to assess it on real world data sets. Firstly, let us define precisely the algorithm.

Algorithm

1. We find the average arithmetic mean and the mean absolute deviation for each variable.
2. We standardize all data set observations by means of subtracting the mean and dividing the difference by the mean absolute deviation. Such standardization is more robust than dividing by the standard deviation (see. Kaufman & Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, 1990).
3. We find the dissimilarity matrix (in all cases it is the Euclidean distance).
4. We find the average aritmetic distance out of a sample 300 distances.
5. We perform the one-step mean shift procedure with the window size found in step 4 to each observation.
6. We find the standard deviation of the values of each variable separately.
7. We rank rank all variables in increasing order with respect to their deviations.

8. The criterion of the greatest jump in deviation decides about the cut of the variables into true and masking ones i.e. we cut at variable *i*-th for which the ratio

$$r(i)=\frac{d(i)-d(i+1)}{d(i-1)-d(i)} \tag{1}$$

where $d(i)$ - deviation of *i*-th variable, is the greatest.

It is worth to note that the most important result is the proper succession of variables.

The criterion given in step 8 is not that important as it may seem. We don't have to break the set of all variables into two subsets, we may use deviations to associate weights with variables and to perform clustering procedures with the set of all variables. We may also use some other criterion.

The problem of assessing the importance of variables in cluster analysis on real world data sets is not so straightforward as it may seem. Some authors (e.g. Dash M, Liu H., (2000)) take into account all variables and use some kind of prior knowledge e.g. implied by associations between variables or experts' opinions about variables, and try to decide which variables are more important than other. In our opinion this kind of approach is dubois – it may happen that all variables are important and in this case we have to use all of them to get proper clustering results. Besides, it is a problem to find experts for every data set to anlyse it properly. We suggest to use the following approach. Add a number of new variables without any cluster pattern e.g. uniformly distributed and try to discover these varaibles as the masking ones. Following Steinley and Brusco (2008) we used a number of variables with the following distributions :

a) Uniform distribution on intreval (0, 20);

b) Normal distribution centered at zero with the covariance matrix equal to the identity matrix;

c) Normal distribution centered at zero with the covariance matrix with ones on the main diagonal and 0.2 out of the diagonal.

The basic difference between our approach and that of Steinley and Brusco (2008) is that we never use the number of masking variables greater than half of all variables. In our opinion it doesn't make any sense in practice to consider more masking variables because if the number of masking variables is too big we do not know any longer which variables are masking and which are true. The number of the masking variables added was equal to 2 for all 4 data sets and the variables' distributions were identical. To compare the performance of our method woth other methods we used the HinoV procedure Carmone et al (1999) with the jump criterion (1) following the idea of Steinley and Brusco (2008).

II. DATA SETS

We used two data sets from UCI Repository and two from the book by Atkinson et al (2004).

The *Iris* data set consists of 150 observations in 4 dimensions and 3 clusters.

The *Pima* data set consists of 200 observations in 7 dimensions and 6 clusters.

The *Diabet* data set consists of 145 observations in 3 dimensions and 3 clusters.

The *Invest* data set consists of 103 observations in 3 dimensions and 2 clusters.

Every data set was converted into a contaminated data set by adding 2 variables with the distributions of type a), b) and c). Thus, altogether, we had 12 data sets to investigate.

III. RESULTS AND CONCLUSIONS

Our method worked very well for all 3 cases of the *Iris* data set – looking at the graphs one doesn't even need any criterion to decide about the choice of variables. The HinoV also worked well.

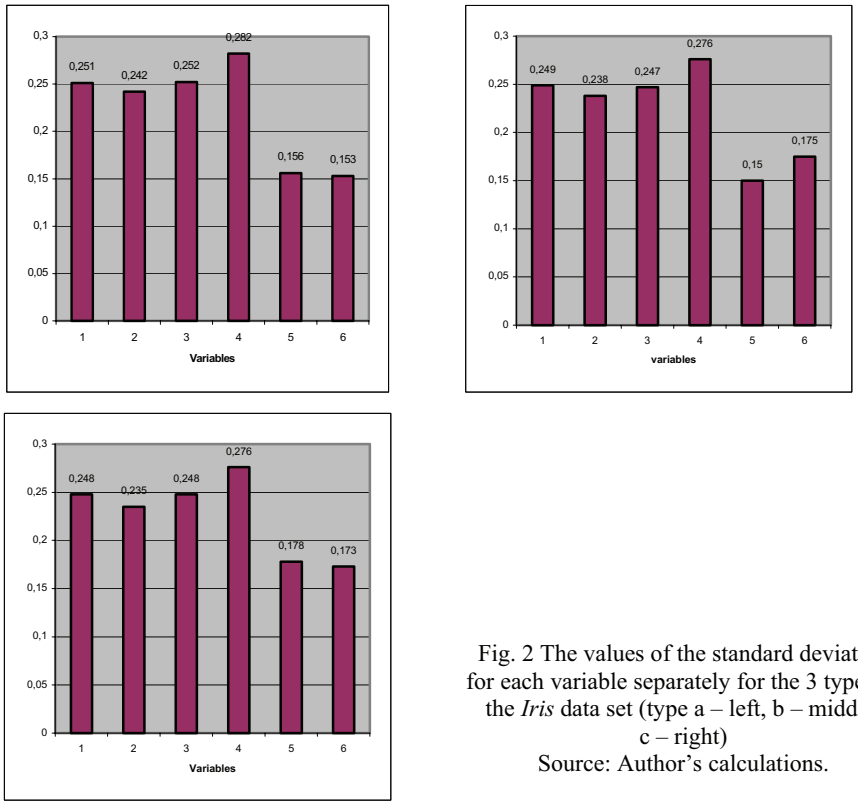


Fig. 2 The values of the standard deviation for each variable separately for the 3 types of the *Iris* data set (type a – left, b – middle, c – right)
Source: Author's calculations.

Our metod worked very well for all 3 cases of the *Daibet* data set – looking at the graphs one doesn't even need any criterion to decide about the choice of variables. The HinoV worked very badly mixing true variables with masking in all three cases.

Our metod worked well for all 3 cases of the *Invest* data set – although, in this first case we had to resort to the criterion because the second variable may seem similar to the last two. The HinoV worked well in all three cases.

Our metod worked well for the second case of the *Pima* data set. For the other two cases it included variable number 6 into the group of true variables. The HinoV worked very badly mixing true variables with masking in all three cases.

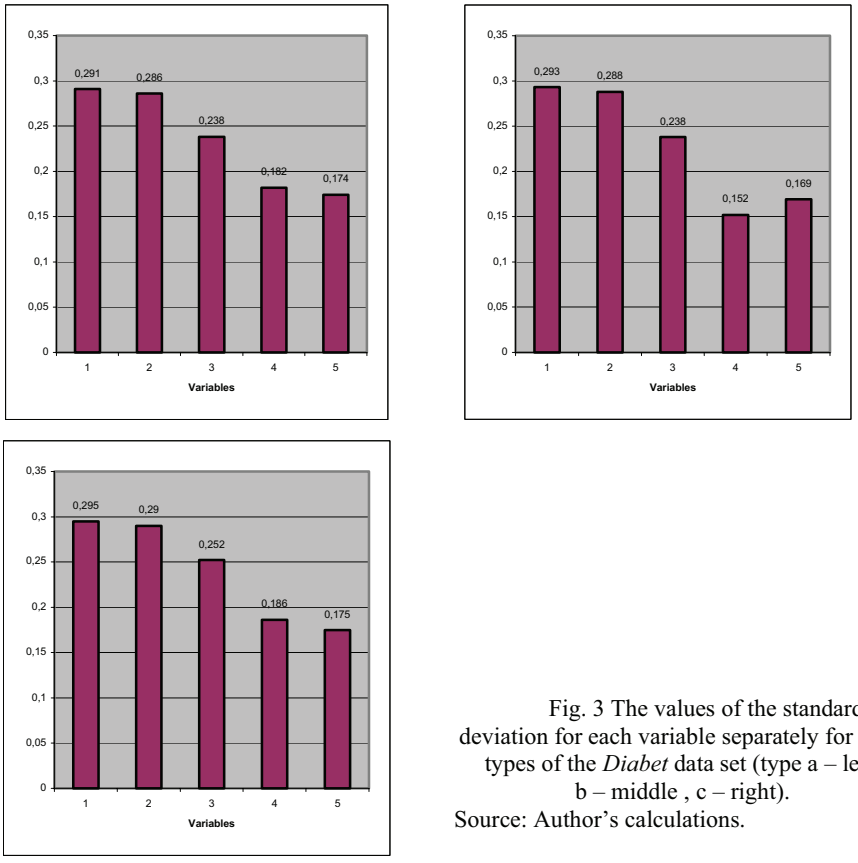
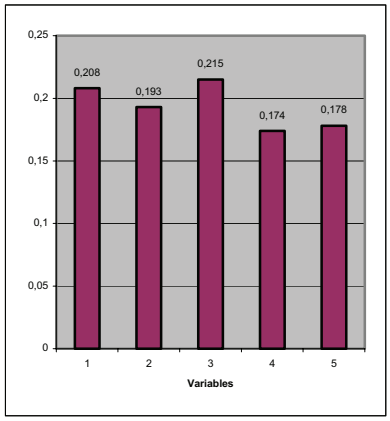
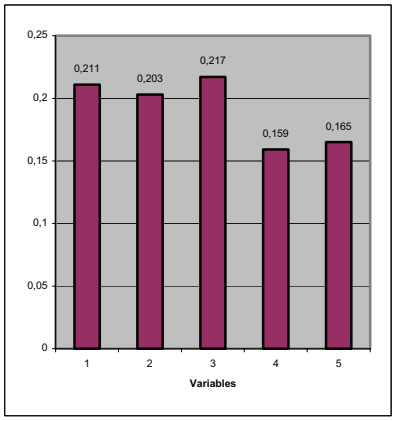


Fig. 3 The values of the standard deviation for each variable separately for the 3 types of the *Diabet* data set (type a – left, b – middle , c – right).
Source: Author's calculations.





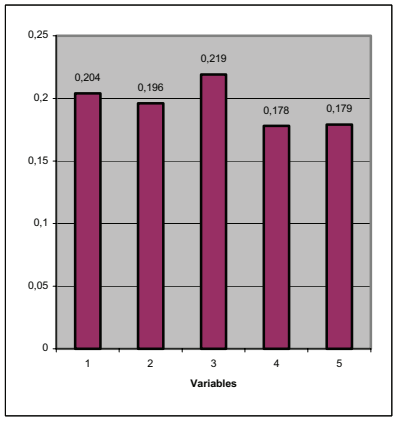
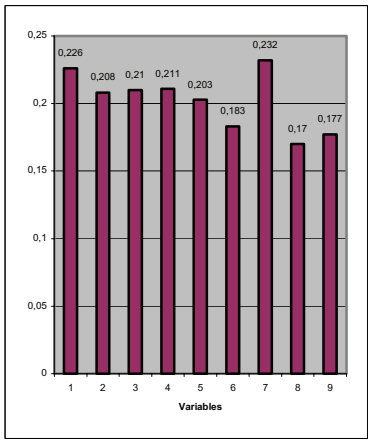
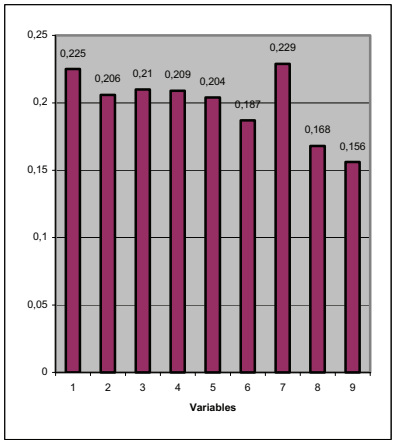


Fig. 4. The values of the standard deviation for each variable separately for the 3 types of the *Invest* data set (type a – left, b – middle, c – right).
Source: Author's calculations





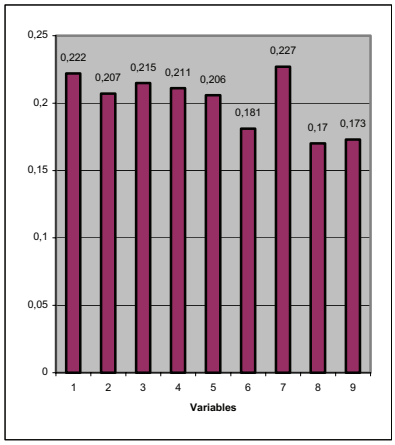


Fig. 5 The values of the standard deviation for each variable separately for the 3 types of the *Pima* data set (type a – left, b – middle, c – right).
Source: Author's calculations.

From the above investigations we may draw the following conclusions.

1. Our method performed better than the compared HinoV method.
2. Our method does not need the specification of the number of clusters – the HinoV does.
3. In the cases of discarding too many variables (cases a) and b) of the *Pima* set) our method does not have too spoil a clustering procedure – we may even get better results. The mistakes done by HinoV were much more troublesome.

REFERENCES

Atkinson A., Riani M. (2004), Cerioli A., *Exploring Multivariate Data with the Forward Search,* Springer-Verlag.

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

Dash M, Liu H., (2000) *Feature selection for clustering* , Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD).

Kaufman L., Rousseeuw P. J., *Finding Groups in Data: an Introduction to Cluster Analysis,* John Wiley&Sons, 1990

Korzeniewski J. (2009), *A novel technique of variable choice and weighting in cluster analysis,* paper at IFCS Drezno 2009.

Steinley D., Brusco M. (2008), *Selection of variables in cluster analysis: an empirical comparison of eight procedures.*, Psychometrika Vol. 73

Jerzy Korzeniewski

BADANIE EFEKTYWNOŚCI NOWEGO ALGORYTMU DO WYBORU ZMIENNYCH W ANALIZIE SKUPIEŃ NA ZBIORACH DANYCH ZE ŚWIATA REALNEGO

W artykule badana jest efektywność algorytmu do wyboru zmiennych w analizie skupień zaproponowanego przez Autora na zbiorach danych ze świata realnego. Ocena tego algorytmu na syntetycznych zbiorach danych w postaci mieszanin rozkładów normalnych była przedmiotem innego badania – algorytm spisał się dobrze. Ideą algorytmu jest wybieranie jako istotnych tych zmiennych, których wariancja nie zmniejsza się tak bardzo jak wariancja zmiennych maskujących po zastosowaniu jednego kroku procedury średniego przesunięcia do wszystkich obserwacji zbioru danych.