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 285, 2013

Jerzy Korzeniewski[*]

MODIFICATION OF TALAVERA METHOD OF VARIABLE SELECTION IN CLUSTER ANALYSIS

Abstract. Talavera has proposed a method of variable selection in cluster analysis for data sets in which only variables measured on nominal scale are present. He examined the method on a couple of data sets basing his assessment on the case in which one can use a data grouping algorithm (he used the COBWEB algorithm). In other approaches some authors try to select variables without referring to any particular grouping method. In the paper, we investigate the efficiency of the Talavera method on real world data sets, referring only to the succession of variables and the greatest jump criterion. Some data sets with variables measured on stronger scales are also investigated after previous descretization.

Key words: cluster analysis, variable choice, correlation of variables.

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 other (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. However, there are few methods designed for data sets with variables measured on weak scales e.g. nominal scale. Talavera (2000) developed a method of variable selection in the context of cluster analysis only for data sets in which there are no continuous variables. In such a case, if we make additional assumption of the existence of no more than one cluster structure, he argues that the variables which are important for the cluster structure should be highly correlated with the rest of true variables. The method was examined in the following way. Firstly, the COBWEB (*Fisher,* 1987) algorithm based on hierarchical tree was applied to group the data. The authors examined two artificial data sets and six from the UCI repository. Every data set was divided randomly into two equal sets – the training

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

set and the test set. Then, the COBWEB algorithm was used to group the training data and the results of the grouping were used to group the test data. The grouping results were assessed by means of a cross-validation test (*Dietterich,* 1998). For the chosen *k* most important variables one can place little credibility in the hypothesis stating that the use of all variables gives smaller percent of false classifications. The numerical measure for choosing *k* was the smallest percent of false classifications (comparing the test and training set). This method is heavily dependant on the use of the COBWEB algorithm, therefore, there is a question of a possibility of omitting this dependence.

The subject of this article is to investigate the efficiency of the modification of the Talavera method focused on excluding the wrapper approach i.e. the dependence on the COBWEB grouping algorithm. In chapter II, a closer description of the Talavera method is given. In the following chapters there is a modification proposal and examination of the efficiency of the modification.

II. TALAVERA METHOD

Correlation between variables is measured in the following way. When a reasonable grouping of the data is given, i.e. the one with homogenous clusters different from one another, then both the fraction of observations from cluster C_k for which variable v assumes value a_{vj} (its j-th variant), which is probability $P\left(C_k \middle| x_v = a_{vj}\right)$, as well as the fraction of the values of variable v which are equal to a_{vj} for observations from cluster C_k which can be written as $P\left(x_v = a_{vj} \middle| C_k\right)$ should be high. In consequence, the quality of the grouping can be measured by means of the number

$$\sum_{k=1}^{K} \sum_{v} \sum_{j} P\left(x_v = a_{vj}\right) P\left(C_k \middle| x_v = a_{vj}\right) P\left(x_v = a_{vj} \middle| C_k\right) \tag{1}$$

In this formula, the symbol $P\left(x_v = a_{vj}\right)$ denotes the fraction of observations a_{vj} among all values of variable v. This symbol (probability) plays the role of the weight ascribed to the product of probabilities deciding about the quality of the grouping. Making use of the Bayes formula we can write

$$P\left(x_v = a_{vj}\right) P\left(C_k \middle| x_v = a_{vj}\right) = P\left(C_k\right) P\left(x_v = a_{vj} \middle| C_k\right) \tag{2}$$

and substituting into (1) we get the following grouping quality measure

$$\sum_{k=1}^{K} P(C_k) \sum_{v} \sum_{j} P^2\left(x_v = a_{vj} \middle| C_k\right) \tag{3}$$

The double inside sum can be treated as the mean number of correctly guessed values of all variables for any object from class C_k . Such interpretation is allowed if we assume that the values of arbitrary variable are guessed with probability $P\left(x_v = a_{vj} \middle| C_k\right)$ and that this value is assumed with the same probability. Under such interpretation the expected number of correctly guessed variables' values without referring to data grouping into clusters is equal to $\sum_{v} \sum_{j} P^2\left(x_v = a_{vj}\right)$. Subtracting this sum from the inside sum of formula (3) we get the increment of the expected number of correctly guessed variables' values that is implied by the knowledge of data grouping into clusters. Formula (3) can be easily used to derive a formula that will describe the dependence of variable v_M on the values of other variables substituting $P\left(x_v = a_{vj}\right)$ for $P(C_k)$ and changing the summation over all *k* clusters C_k for the summation over all variables v and their variants *j*. Making use of the subtraction of probabilities and averaging the result we get expression

$$Corr\left(v_M\right) = \frac{\sum_{v} \sum_{j} P\left(x_v = a_{vj}\right) \sum_{j_M} \left(P^2\left(x_{v_M} = a_{vj_M} \middle| x_v = a_{vj}\right) - P^2\left(x_{v_M} = a_{vj_M}\right)\right)}{\left|\{v | v \neq v_M\}\right|} \tag{4}$$

where the quantity in the denominator is the number of attributes different from attribute v_M .

The expression (4) can be used to arrange all variables in order from the variable which is most correlated with other variables to the one that is least correlated. When all attributes are measured on nominal scale such order of variables can be thought of as equivalent with the hierarchy of variables importance for the data set structure. Once the variables are ordered we can use this ordering to pick up first *k* variables and reject the rest. However, the breaking of the sequence of all variables into two parts is dependent on the grouping algorithm and the number of clusters which has to be known. Talavera used the COBWEB algorithm to group the training data and the results of the grouping were used to group the test data (test data and training data are roughly the same size and pick

up randomly). The grouping results were assessed by means of a cross-validation test (*Dietterich,* 1998). The number *k* was determined by the smallest percentage of classification mistakes.

III. MODIFICATION OF TALAVERA METHOD

It seems interesting to check if we actually need to know the number of clusters and base our results on a grouping algorithm. These two obstacles are very troublesome in practical data set considerations. Therefore, the following modification of the Talavera method was proposed.

Step 1 Arrange all variables in decreasing order of their importance to the data set cluster structure measured with correlation given by formula (4).

Step 2 Find the "elbow" on the graph of the correlations (4), similarly as in the HINoV procedure. Pick up the attributes before the elbow as important for the cluster structure and reject the rest.

In practical data set applications we can inspect the graph for the elbow visually. If it is not clearly visible (as well as in simulation experiments) we have to resort to the greatest jump criterion i.e. we choose first *k* attributes for which the relation of the increment of correlation to the increment of correlation for the first *k*+1 is greatest.

IV. INVESTIGATION OF THE MODIFICATION'S EFFICIENCY

Large simulation experiments are not used widely for nominal attributes, probably, due to the far reaching arbitrariness of defining cluster structures. We investigated the efficiency of the modification on a couple of data sets from the UCI repository. We included sets with continuous variables after subjecting them to previous descretization procedure. The descretization consisted in dividing the marginal histogram of each variable into 5 bins of equal width and assigning labels to the observations accordingly.

The research was organized as follows. Firstly, we decided on the number of noisy attributes that should be added to the original variables. Usually, we considered one case of roughly equal number of relevant and irrelevant attributes. In some cases, when adding this number of variables gave very poor performance, we also tried a smaller number of noisy variables. Then we had to choose the type of the distribution for the noisy attributes. As correlated attributes are not allowed in this case (they would create second cluster structure after descretization), we decided to use only uncorrelated attributes: the standardized normal, the uniform distribution on interval [0, 30] and beta(1,1). From the pooled set of original and noisy attributes we were choosing a number of attributes which were considered to create the cluster structure. Results are presented below.

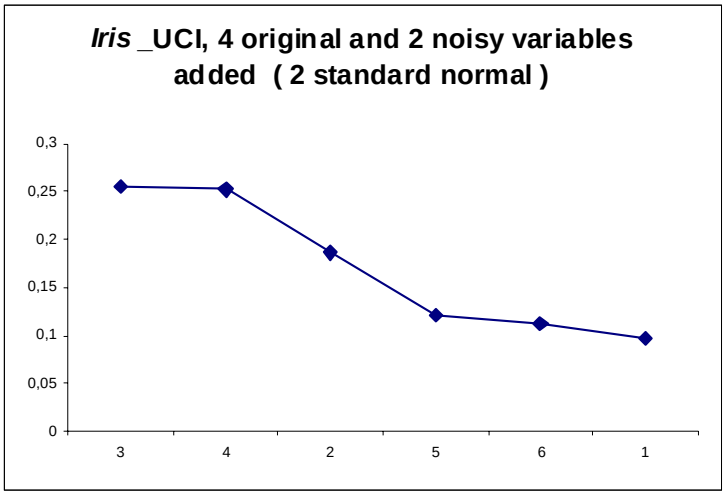

Figure 1. Correlations of single variables with the rest of the variables f or the *Iris*_UCI data set and 2 noisy variables added
Source: own work.

Set 1. *Iris*_UCI data set. Objects: 150. Original variables: 4 continuous variables. Noisy variables added: 2 standard normal variables. Quite good performance, elbow clearly visible, only one variable (variable number 1) lost (compare Fig.1).

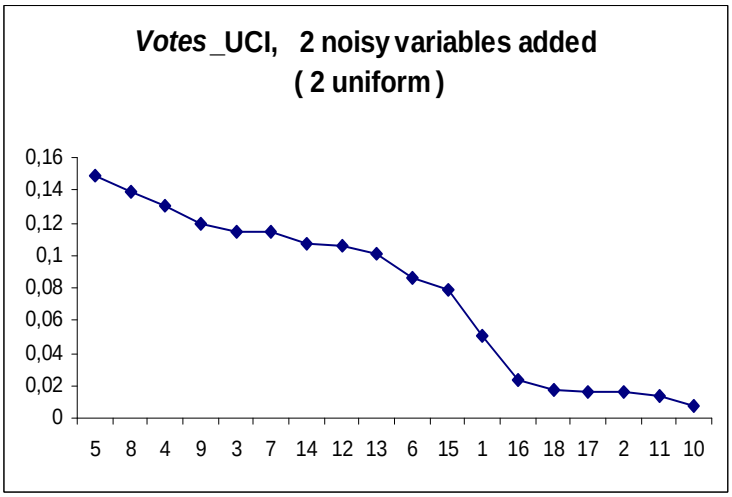

Figure 2. Correlations of single variables with the rest of the variables for the *Votes*_UCI data set and 2 noisy variables added.
Source: own work.

Set 2. *Votes*_UCI data set. Objects: 435. Original variables: 16 nominal variables. Noisy variables added: case a) 2 standard normal, case b) 3 standard normal, 3 uniform and 3 beta variables. Rather poor performance (compare Fig. 2 and Fig. 3), elbow clearly visible but some true variables rejected along with 2 noisy variables In case b) almost all noisy variables were accepted as true. A comment is necessary in this place because this data set is known to contain noisy variables (*Talavera,* 2000). Thus, if e.g. variables 2, 10, 11 were not important for the cluster structure, the result in case a) should be considered to be very good.

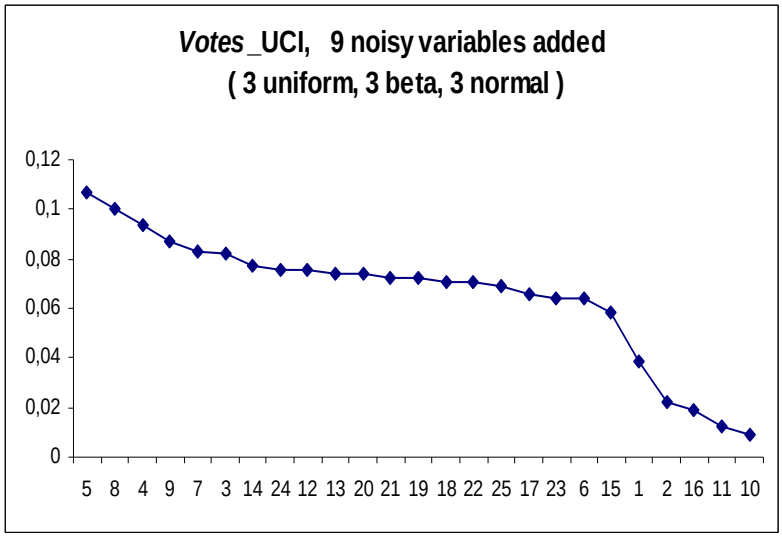

Figure 3. Correlations of single variables with the rest of the variables for the *Votes*_UCI data set and 9 noisy variables added.
Source: own work.

Set 3. *Teaching*_UCI data set. Objects: 151. Original variables: 5 ordinal variables.

Noisy variables added: case a) 2 standard normal, case b) 2 uniform. Very poor performance in both cases (compare Fig. 4 and Fig. 5), elbow invisible and noisy variables at the beginning.

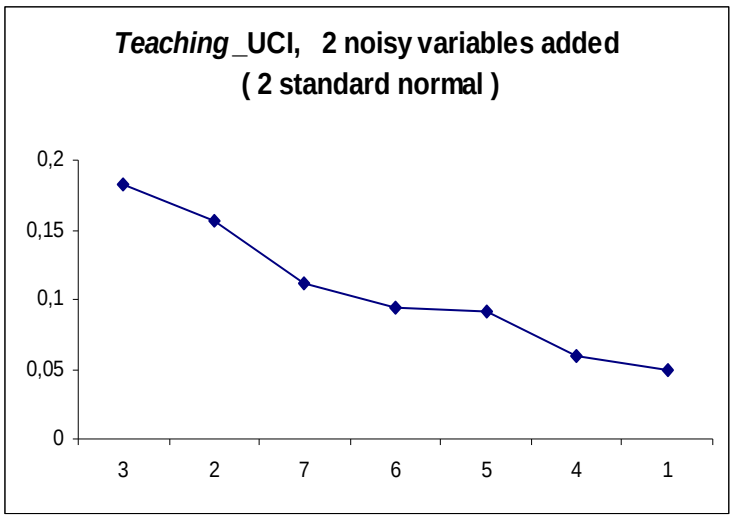

Figure 4. Correlations of single variables with the rest of the variables for the *Teaching*_UCI data set and 2 noisy variables added.
Source: own work.

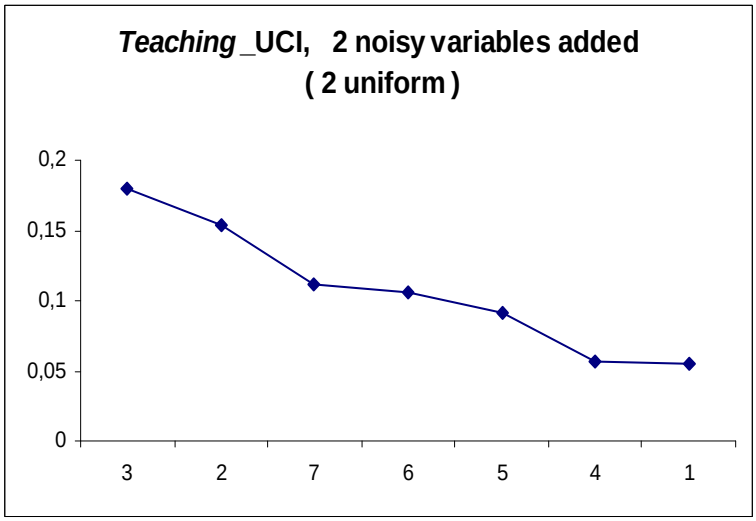

Figure 5. Correlations of single variables with the rest of the variables for the *Teaching*_UCI data set and 2 noisy variables added.
Source: own work.

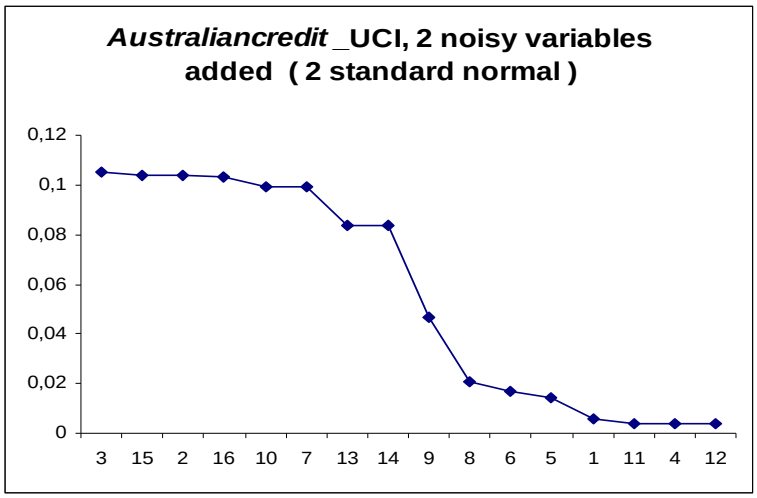

Figure 6. Correlations of single variables with the rest of the variables for the *Australiancredit*_UCI data set and 2 noisy variables added.
Source: own work.

Set 4. *Australiencredit*_UCI data set. Objects: 690. Original variables: 4 nominal, 2 ordinal, 8 continuous. Noisy variables added: 2 standard normal. Very poor performance (compare Fig. 6), in spite of a very small number of noisy variables, elbow visible, but both noisy variables are included at the very beginning.

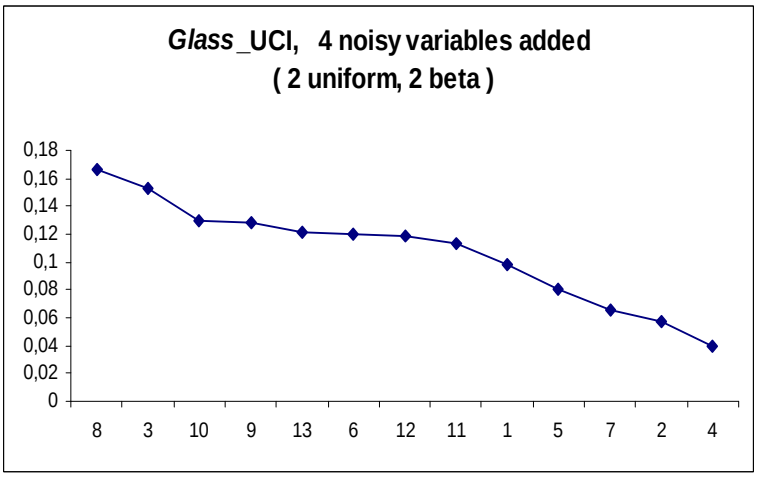

Figure 7. Correlations of single variables with the rest of the variables for the *Glass*_UCI data set and 4 noisy variables added.
Source: own work.

Set 5. *Glass*_UCI data set. Objects: 214. Original variables: 9 continuous variables. Noisy variables added: 2 uniform, 2 beta. Very poor performance (compare Fig. 7), elbow invisible and some noisy variables at the beginning.

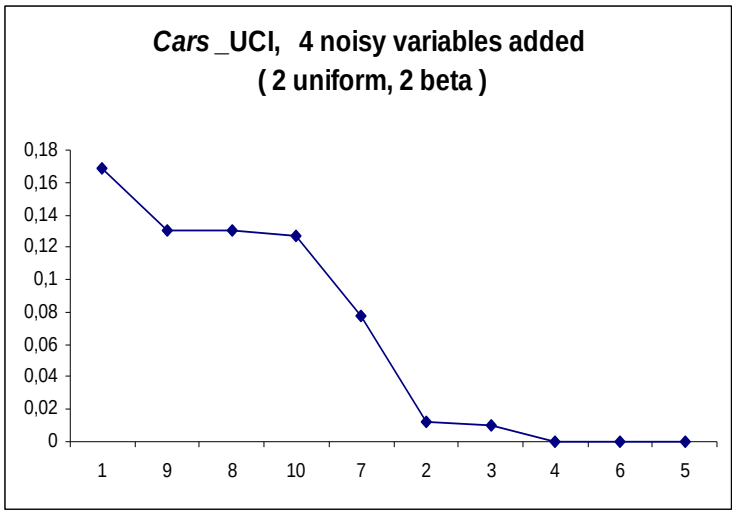

Figure 8. Correlations of single variables with the rest of the variables for the *Cars*_UCI data set and 4 noisy variables added.
Source: own work.

Set 6. *Cars*_UCI data set. Objects: 1728. Original variables: 6 ordinal variables. Noisy variables added: 2 uniform, 2 beta. Very poor performance (compare Fig. 8), elbow clearly visible but all noisy variables at the beginning.

V. RESULTS AND CONCLUSIONS

The instances of the real world data sets investigated allow to formulate the following conclusions.

- The Talavera method has restricted applicability because only one cluster structure is allowed, noisy variables cannot be correlated, most of the features must not be continuous.
- The real world data sets investigated suggest that we cannot replace the original method of Talavera with the modification proposed based on the visual assessment of the correlation graph - the frequency of wrong decisions or inabilities to make any decision was too high.

- However, the order of variables resulting from the correlation analysis is very often incorrect, so, even the use of any grouping procedure will not give proper result.

It is worthwhile to observe that the last conclusion questions the sense of the original form of the Talavera method, because, if the order of variables is incorrect the final choice of variables cannot be correct. This conclusion is limited though to the case of correlated variables being present among the noisy variables.

REFERENCES

Dietterich, T. G., (1998). *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, Neural Computation, 10.

Fisher D., (1987), *Knowledge acquisition via incremental conceptual clustering,* PhD. Thesis, University of California, Irvine.

Talavera L., (2000), *Dependency-Based Feature Selection for Clustering Symbolic Data*, Intelligent Data Analysis 4.

Jerzy Korzeniewski

BADANIE EFEKTYWNOŚCI MODYFIKACJI METODY TALAVERY WYBIERANIA ZMIENNYCH W ANALIZIE SKUPIEŃ NA EMPIRYCZNYCH ZBIORACH DANYCH

Talavera zaproponował metodę wybierania zmiennych tworzących strukturę skupień w zbiorze danych dla zbiorów, w których występują tylko zmienne mierzone na skali nominalnej. Autor zbadał tę metodę na kilku empirycznych zbiorach opierając ocenę na tym jak spisywała się metoda w połączeniu z ustalonym sposobem grupowania danych (algorytm COBWEB). W innych podejściach do tego samego zagadnienia autorzy starają się oprzeć wybór zmiennych na samym uporządkowaniu zbioru zmiennych bez odwoływania się do grupowania obserwacji. W artykule badana jest efektywność metody również w odniesieniu do empirycznych zbiorów danych, uzależniona tylko od uporządkowania zmiennych, oparta na kryterium największego skoku. Rozważane są również zbiory z niektórymi zmiennymi mierzonymi na mocniejszych skalach z po uprzedniej dyskretyzacji zmiennych.