Identification of noisy variables for nonmetric and symbolic data in cluster analysis

Marek Walesiak and Andrzej Dudek

Wroclaw University of Economics, Department of Econometrics and Computer Science, Nowowiejska 3, 58-500 Jelenia Gora, Poland
marek.walesiak@ae.jgora.pl, andrzej.dudek@ae.jgora.pl

Abstract. A proposal of an extended version of the HINoV method for the identification of the noisy variables (Carmone et al [1999]) for nonmetric, mixed, and symbolic interval data is presented in this paper. Proposed modifications are evaluated on simulated data from a variety of models. The models contain the known structure of clusters. In addition, the models contain a different number of noisy (irrelevant) variables added to obscure the underlying structure to be recovered.

1 Introduction

Choosing variables is the one of the most important steps in a cluster analysis. Variables used in applied clustering should be selected and weighted carefully. In a cluster analysis we should include only those variables that are believed to help to discriminate the data (see Milligan [1996], p. 348). Two classes of approaches, while choosing the variables for cluster analysis, can facilitate a cluster recovery in the data (see e.g. Gnanadesikan et al [1995]; Milligan [1996], pp. 347-352):

– variable selection (selecting a subset of relevant variables),

– variable weighting (introducing relative importance of the variables according to their weights).

Carmone et al [1999] discussed the literature on the variable selection and weighting (the characteristics of six methods and their limitations) and proposed the HINoV method for the identification of the noisy variables, in the area of the variable selection, to remedy problems with these methods. They demonstrated its robustness with metric data and k-means algorithm. The authors suggest further studies of the HINoV method with different types of data and other clustering algorithms on p. 508.

In this paper we propose extended version of the HINoV method for nonmetric, mixed, and symbolic interval data. The proposed modifications are evaluated for eight clustering algorithms on simulated data from a variety of models.

2 Characteristics of the HINoV method and its modifications

Algorithm of Heuristic Identification of Noisy Variables (HINoV) method for metric data (see Carmone et al [1999]) is following:

1. A data matrix $[x_{ij}]$ containing n objects and m normalized variables measured on a metric scale ($i = 1, \ldots, n$; $j = 1, \ldots, m$) is a starting point.

2. Cluster, via kmeans method, the observed data separately for each j-th variable for a given number of clusters u. It is possible to use clustering methods based on a distance matrix (pam or any hierarchical agglomerative method: single, complete, average, mcquitty, median, centroid, Ward).

3. Calculate adjusted Rand indices R_{jl} ($j, l = 1, \ldots, m$) for partitions formed from all distinct pairs of the m variables ($j \neq l$). Due to a fact that adjusted Rand index is symmetrical we need to calculate $m(m-1)/2$ values.

4. Construct $m \times m$ adjusted Rand matrix (parim). Sum rows or columns for each j-th variable $R_{j\bullet} = \sum_{l=1}^{m} R_{jl}$ (topri):

$$\begin{matrix} \text{Variable} \\ \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_m \end{bmatrix} \end{matrix} \quad \begin{matrix} \text{parim} \\ \begin{bmatrix} & R_{12} & \ldots & R_{1m} \\ R_{21} & & \ldots & R_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ R_{m1} & R_{m2} & \ldots & \end{bmatrix} \end{matrix} \quad \begin{matrix} \text{topri} \\ \begin{bmatrix} R_{1\bullet} \\ R_{2\bullet} \\ \vdots \\ R_{m\bullet} \end{bmatrix} \end{matrix}$$

5. Rank topri values $R_{1\bullet}, R_{2\bullet}, \ldots, R_{m\bullet}$ in a decreasing order (stopri) and plot the scree diagram. The size of the topri values indicate a contribution of that variable to the cluster structure. A scree diagram identifies sharp changes in the topri values. Relatively low-valued topri variables (the noisy variables) are identified and eliminated from the further analysis (say h variables).

6. Run a cluster analysis (based on the same classification method) with the selected $m - h$ variables.

The modification of the HINoV method for nonmetric data (where number of objects is much more than a number of categories) differs in steps 1, 2, and 6 (see Walesiak [2005]):

1. A data matrix $[x_{ij}]$ containing n objects and m ordinal and/or nominal variables is a starting point.

2. For each j-th variable we receive natural clusters, where the number of clusters equals the number of categories for that variable (for instance five for Likert scale or seven for semantic differential scale).

6. Run a cluster analysis with one of clustering methods based on a distance appropriate to nonmetric data (GDM2 for ordinal data – see Jajuga et al [2003]; Sokal and Michener distance for nominal data) with the selected $m-h$ variables.

The modification of the HINoV method for symbolic interval data differs in steps 1 and 2:

1. A symbolic data array containing n objects and m symbolic interval variables is a starting point.

2. Cluster the observed data with one of clustering methods (pam, single, complete, average, mcquitty, median, centroid, Ward) based on a distance appropriate to the symbolic interval data (e.g. Hausdorff distance - see Billard and Diday [2006], p. 246) separately for each j-th variable for a given number of clusters u.

Functions HINoV.Mod and HINoV.Symbolic of clusterSim computer program working in R allow adequately using mixed (metric, nonmetric), and the symbolic interval data. The proposed modifications of the HINoV method are evaluated on simulated data from a variety of models.

3 Simulation models

We generate data sets in eleven different scenarios. The models contain the known structure of clusters. In the models 2-11 the noisy variables are simulated independently from the uniform distribution.

Model 1. No cluster structure. 200 observations are simulated from the uniform distribution over the unit hypercube in 10 dimensions (see Tibshirani et al [2001], p. 418).

Model 2. Two elongated clusters in 5 dimensions (3 noisy variables). Each cluster contains 50 observations. The observations in each of the two clusters are independent bivariate normal random variables with means (0, 0), (1, 5), and covariance matrix $\sum$ ($\sigma_{jj} = 1$, $\sigma_{jl} = -0.9$).

Model 3. Three elongated clusters in 7 dimensions (5 noisy variables). Each cluster is randomly chosen to have 60, 30, 30 observations, and the observations are independently drawn from bivariate normal distribution with means (0, 0), (1.5, 7), (3, 14) and covariance matrix $\sum$ ($\sigma_{jj} = 1$, $\sigma_{jl} = -0.9$).

Model 4. Three elongated clusters in 10 dimensions (7 noisy variables). Each cluster is randomly chosen to have 70, 35, 35 observations, and the observations are independently drawn from multivariate normal distribution with means (1.5, 6, –3), (3, 12, –6), (4.5, 18, –9), and identity covariance matrix $\sum$, where $\sigma_{jj} = 1$ ($1 \leq j \leq 3$), $\sigma_{12} = \sigma_{13} = -0.9$, and $\sigma_{23} = 0.9$.

Model 5. Five clusters in 3 dimensions that are not well separated (1 noisy variable). Each cluster contains 25 observations. The observations are independently drawn from bivariate normal distribution with means (5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5), and identity covariance matrix $\sum$ ($\sigma_{jj} = 1$, $\sigma_{jl} = 0.9$).

Model 6. Five clusters in 5 dimensions that are not well separated (2 noisy variables). Each cluster contains 30 observations. The observations are independently drawn from multivariate normal distribution with means (5, 5, 5), (–3, 3, –3), (3, –3, 3), (0, 0, 0), (–5, –5, –5), and covariance matrix $\sum$, where $\sigma_{jj} = 1$ ($1 \leq j \leq 3$), and $\sigma_{jl} = 0.9$ ($1 \leq j \neq l \leq 3$).

Model 7. Five clusters in 10 dimensions (8 noisy variables). Each cluster is randomly chosen to have 50, 20, 20, 20, 20 observations, and the observations are independently drawn from bivariate normal distribution with means (0, 0), (0, 10), (5, 5), (10, 0), (10, 10), and identity covariance matrix $\sum$ ($\sigma_{jj} = 1$, $\sigma_{jl} = 0$).

Model 8. Five clusters in 9 dimensions (6 noisy variables). Each cluster contains 30 observations. The observations are independently drawn from multivariate normal distribution with means (0, 0, 0), (10, 10, 10), (–10, –10, –10), (10, –10, 10), (–10, 10, 10), and identity covariance matrix $\sum$, where $\sigma_{jj} = 3$ $(1 \leq j \leq 3)$, and $\sigma_{jl} = 2$ $(1 \leq j \neq l \leq 3)$.

Model 9. Four clusters in 6 dimensions (4 noisy variables). Each cluster is randomly chosen to have 50, 50, 25, 25 observations, and the observations are independently drawn from bivariate normal distribution with means (–4, 5), (5, 14), (14, 5), (5, –4), and identity covariance matrix $\sum$ ($\sigma_{jj} = 1$, $\sigma_{jl} = 0$).

Model 10. Four clusters in 12 dimensions (9 noisy variables). Each cluster contains 30 observations. The observations are independently drawn from multivariate normal distribution with means (–4, 5, –4), (5, 14, 5), (14, 5, 14), (5, –4, 5), and identity covariance matrix $\sum$, where $\sigma_{jj} = 1$ $(1 \leq j \leq 3)$, and $\sigma_{jl} = 0$ $(1 \leq j \neq l \leq 3)$.

Model 11. Four clusters in 10 dimensions (9 noisy variables). Each cluster contains 35 observations. The observations on the first variable are independently drawn from univariate normal distribution with means –2, 4, 10, 16 respectively, and identity variance $\sigma_j^2 = 0.5$ $(1 \leq j \leq 4)$.

Ordinal data. The clusters in models 1-11 contain continuous data and a discretization process is performed on each variable to obtain ordinal data. The number of categories k determines the width of each class intervals: $\left[\max_i\{x_{ij}) - \min_i\{x_{ij}\}\right] \Big/ k$. Independently for each variable each class interval receive category $1, \ldots, k$ and the actual value of variable x_{ij} is replaced by these categories. In simulation study $k = 5$ (for $k = 7$ we have received similar results).

Symbolic interval data. To obtain symbolic interval data the data were generated for each model twice into sets A and B and minimal (maximal) value of $\{a_{ij}, b_{ij}\}$ is treated as the beginning (the end) of an interval.

Fifty realizations were generated from each setting.

4 Discussion on the simulation results

In testing the robustness of the HINoV modified algorithm using simulated ordinal or symbolic interval data, the major criterion was the identification of the noisy variables. The HINoV-selected variables contain variables with the highest topri values. In models 2-11 the number of nonnoisy variables is known. Due to this fact, in simulation study, the number of the HINoV-selected variables equals the number of nonnoisy variables in each model.

When the noisy variables were identified, the next step was to run the one of clustering methods based on distance matrix (pam, single, complete, average, mcquitty, median, centroid, Ward) with the nonnoisy subset of variables (HINoV-selected variables) and with all variables. Then each clustering result was compared with the known cluster structure from models 2-11 using Hubert and Arabie's [1985] corrected Rand index (see Table 1 and 2).

Table 1. Cluster recovery for all variables and HINoV-selected subsets of variables for ordinal data (five categories) by experimental model and clustering method

Model		Clustering method							
		pam	ward	single	complete	average	mcquitty	median	centroid
2	a	0.38047	0.53576	0.00022	0.11912	0.42288	0.25114	0.00527	0.00032
	b	0.84218	0.90705	0.72206	0.12010	0.99680	0.41796	0.30451	0.89835
3	a	0.27681	0.34071	0.00288	0.29392	0.40818	0.35435	0.04625	0.00192
	b	0.85946	0.60606	0.36121	0.61090	0.68223	0.51487	0.49199	0.61156
4	a	0.35609	0.44997	0.00127	0.43860	0.53509	0.47083	0.04677	0.00295
	b	0.83993	0.87224	0.56313	0.56541	0.80149	0.62102	0.54109	0.80156
5	a	0.54746	0.60139	0.27610	0.46735	0.58050	0.49842	0.33303	0.50178
	b	0.91071	0.84888	0.48550	0.73720	0.81317	0.79644	0.72899	0.74462
6	a	0.61074	0.60821	0.13400	0.53296	0.61037	0.56426	0.35113	0.47885
	b	0.83880	0.87183	0.56074	0.75584	0.86282	0.81395	0.71085	0.79018
7	a	0.10848	0.11946	0.00517	0.09267	0.10945	0.11883	0.00389	0.00659
	b	0.80072	0.87399	0.27965	0.87892	0.94882	0.77503	0.74141	0.91638
8	a	0.31419	0.43180	0.00026	0.29529	0.40203	0.36771	0.00974	0.00023
	b	0.95261	0.96372	0.58026	0.95596	0.96627	0.95507	0.93701	0.96582
9	a	0.37078	0.45915	0.01123	0.12128	0.50198	0.31134	0.04326	0.00709
	b	0.99966	0.98498	0.93077	0.96993	0.99626	0.98024	0.95461	0.99703
10	a	0.29727	0.41152	0.00020	0.22358	0.41107	0.34663	0.00030	0.00007
	b	1.00000	1.00000	0.99396	0.99911	1.00000	1.00000	0.99867	1.00000
	$\bar{b}$	0.89378	0.88097	0.60858	0.73259	0.89642	0.76384	0.71212	0.85838
	$\bar{r}$	0.53130	0.44119	0.56066	0.44540	0.45403	0.39900	0.61883	0.74730
ccr		98.22%	98.00%	94.44%	90.67%	97.11%	89.56%	98.89%	98.44%
11	a	0.04335	0.04394	0.00012	0.04388	0.03978	0,03106	0,00036	0.00009
	b	0.14320	0.08223	0.12471	0.08497	0.10373	0,12355	0,04626	0,06419

a (b) – values represent Hubert and Arabie's adjusted Rand indices averaged over fifty replications for each model with all variables (with HINoV-selected variables); $\bar{r} = \bar{b} - \bar{a}$; ccr – corrected cluster recovery.

Some conclusions can be drawn from the simulations results:

1. The cluster recovery that used only the HINoV-selected variables for ordinal data (Table 1) and symbolic interval data (Table 2) was better than the one that used all variables for all models 2-10 and each clustering method.

2. Among 450 simulated data sets (nine models with 50 runs) the HINoV method was better (see ccr in Table 1 and 2):

Table 2. Cluster recovery for all variables and HINoV-selected subsets of variables for symbolic interval data by experimental model and clustering method

Model		Clustering method							
		pam	ward	single	complete	average	mcquitty	median	centroid
2	a	0.86670	0.87920	0.08006	0.28578	0.32479	0.49424	0.02107	0.00004
	b	0.99920	0.97987	0.91681	0.99680	0.99524	0.98039	0.85840	0.95739
3	a	0.41934	0.39743	0.00368	0.37361	0.38831	0.36597	0.00088	0.00476
	b	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99062	1.00000
4	a	0.04896	0.01641	0.00269	0.01653	–0.00075	0.01009	0.00177	0.00023
	b	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
5	a	0.71543	0.70144	0.73792	0.47491	0.60960	0.53842	0.34231	0.28338
	b	0.99556	0.99718	0.98270	0.91522	0.99478	0.99210	0.90252	0.97237
6	a	0.75308	0.67237	0.33392	0.47230	0.67817	0.55727	0.18194	0.10131
	b	0.99631	0.99764	0.99169	0.95100	0.98809	0.97881	0.84463	0.99866
7	a	0.36466	0.51262	0.00992	0.32856	0.33905	0.39823	0.00527	0.00681
	b	1.00000	0.99974	1.00000	0.98493	0.99954	1.00000	0.99974	0.99954
8	a	0.74711	0.85104	0.01675	0.50459	0.51029	0.61615	0.00056	0.00023
	b	1.00000	0.99966	0.99932	0.99966	0.99966	0.99843	0.99835	1.00000
9	a	0.86040	0.90306	0.30121	0.26791	0.54639	0.62620	0.00245	0.00419
	b	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
10	a	0.70324	0.91460	0.00941	0.48929	0.47886	0.54275	0.00007	0.00004
	b	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	$\bar{b}$	0.99900	0.99712	0.98783	0.98306	0.99747	0.99441	0.95491	0.99199
	$\bar{r}$	0.39023	0.34732	0.82166	0.62601	0.56687	0.53337	0.89310	0.94744
ccr		94.67%	91.78%	97.33%	99.11%	96.22%	96.44%	99.56%	99.78%
11	a	0.05334	0.04188	0.00007	0.03389	0.02904	0.03313	0.00009	0.00004
	b	0.12282	0.04339	0.04590	0.08259	0.08427	0.14440	0.04380	0.08438

a (b); $\bar{r} = \bar{b} - \bar{a}$; *ccr* – see Table 1.

– from 89.56% (mcquitty) to 98.89% (median) of runs for ordinal data,

– from 91.78% (ward) to 99,78% (centroid) of runs for symbolic interval data.

3. Figure 1 shows the relationship between the values of adjusted Rand indices averaged over fifty replications and models 2-10 with the HINoV-selected variables ($\bar{b}$) and values showing an improvement ($\bar{r}$) of average adjusted Rand indices (cluster recovery with the HINoV selected variables against all variables) separately for eight clustering methods and types of data (ordinal, symbolic interval). Based on adjusted Rand indices averaged over fifty replications and models 2-10 the improvements in cluster recovery (HINoV selected variables against all variables) are varying:

– for ordinal data from 0.3990 (mcquitty) to 0.7473 (centroid),

– for symbolic interval data from 0.3473 (ward) to 0.9474 (centroid).

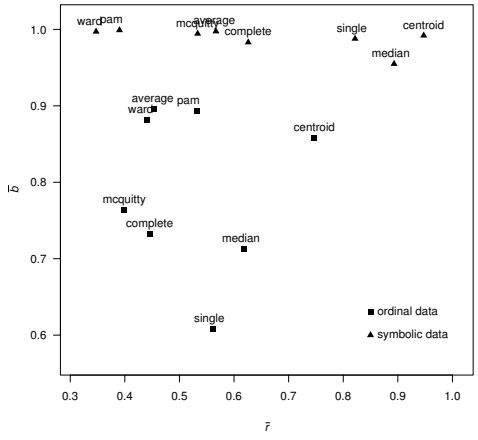

Fig. 1. The relationship between values of $\bar{b}$ and $\bar{r}$
Source: own research

5 Conclusions

The HINoV algorithm has limitations for analyzing nonmetric and symbolic interval data almost the same as the ones mentioned in Carmone et al [1999] article for metric data.

First, the HINoV is of a little use with a nonmetric data set or a symbolic data array in which all variables are noisy (no cluster structure – see model 1). In this situation topri values are similar and close to zero (see Table 3).

Table 3. Mean and standard deviation of topri values for 10 variables in model 1

Variable	Ordinal data with five categories		Symbolic data array	
	mean	sd	mean	sd
1	–0.00393	0.01627	0.00080	0.02090
2	–0.00175	0.01736	0.00322	0.02154
3	0.00082	0.02009	0.00179	0.01740
4	–0.00115	0.01890	–0.00206	0.02243
5	0.00214	0.02297	–0.00025	0.02074
6	0.00690	0.02030	–0.00312	0.02108
7	–0.00002	0.02253	–0.00440	0.02044
8	0.00106	0.01754	0.00359	0.01994
9	0.00442	0.01998	0.00394	0.02617
10	–0.00363	0.01959	0.00023	0.02152

Second, the HINoV method depends on the relationship between pairs of variables. If we have only one variable with a cluster structure and the others are noisy, the HINoV will not be able to isolate this nonnoisy variable (see Table 4).

Table 4. Mean and standard deviation of topri values for 10 variables in model 11

Variable	Ordinal data with five categories		Symbolic data array	
	mean	sd	mean	sd
1	–0.00095	0.03050	0.00012	0.02961
2	–0.00198	0.02891	0.00070	0.03243
3	0.00078	0.02937	–0.00206	0.02969
4	–0.00155	0.02950	–0.00070	0.03185
5	0.00056	0.02997	–0.00152	0.03157
6	0.00148	0.03090	–0.00114	0.03064
7	–0.00246	0.02959	–0.00203	0.03019
8	–0.00274	0.03137	–0.00186	0.03021
9	–0.00099	0.02975	0.00088	0.03270
10	0.00023	0.02809	–0.00181	0.03126

Third, if all variables have the same cluster structure (no noisy variables) the topri values will be large and similar for all variables. The suggested selection process using a scree diagram will be ineffective.

Fourth, an important problem is to decide on a proper number of clusters in stage two of the HINoV algorithm with symbolic interval data. To resolve this problem we should initiate the HINoV algorithm with a different number of clusters.

References

BILLARD, L., DIDAY, E. (2006): *Symbolic data analysis. Conceptual statistics and data mining*, Wiley, Chichester.

CARMONE, F.J., KARA, A. and MAXWELL, S. (1999): *HINoV: a new method to improve market segment definition by identifying noisy variables*, "Journal of Marketing Research", vol. 36, November, 501-509.

GNANADESIKAN, R., KETTENRING, J.R., and TSAO, S.L. (1995): *Weighting and selection of variables for cluster analysis*, "Journal of Classification", vol. 12, no. 1, 113-136.

HUBERT, L.J., ARABIE, P. (1985): *Comparing partitions*, "Journal of Classification", vol. 2, no. 1, 193-218.

JAJUGA, K., WALESIAK, M., BAK, A. (2003): *On the General Distance Measure*, In: M., Schwaiger, and O., Opitz (Eds.), *Exploratory data analysis in empirical research*, Springer-Verlag, Berlin, Heidelberg, 104-109.

MILLIGAN, G.W. (1996): *Clustering validation: results and implications for applied analyses*, In: P., Arabie, L.J., Hubert, G., de Soete (Eds.), *Clustering and classification*, World Scientific, Singapore, 341-375.

TIBSHIRANI, R., WALTHER, G., HASTIE, T. (2001): *Estimating the number of clusters in a data set via the gap statistic*, "Journal of the Royal Statistical Society", ser. B, vol. 63, part 2, 411-423.

WALESIAK, M. (2005): *Variable selection for cluster analysis – approaches, problems, methods*, Plenary Session of the Committee on Statistics and Econometrics of the Polish Academy of Sciences, 15, March, Wroclaw.