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, 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 (Milligan (1996), p. 348). Two classes of approaches, while choosing the variables for cluster analysis, can facilitate a cluster recovery in the data (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 (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):

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

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 (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.