Cluster Quality Indexes for Symbolic Classification – An Examination

Andrzej Dudek

Wrocław University of Economics, Department of Econometrics and Computer Science, Nowowiejska 3, 58-500 Jelenia Góra, Poland; andrzej.dudek@ae.jgora.pl

Abstract. The paper presents difficulties of measuring clustering quality for symbolic data (such as lack of a "traditional" data matrix). Some hints concerning the usage of known indexes for such kind of data are given and indexes designed exclusively for symbolic data are described. Finally, after the presentation of simulation results, some proposals for choosing the most adequate indexes for popular classification algorithms are given.

1 Introduction

In a typical classification procedure, cluster validation is one of the crucial steps. Typically, in the validation step an internal cluster quality index is used. There is a variety of such kind of indexes with over fifty measures (Milligan and Cooper (1985), Weingessel et al. (1999))

The problem of choosing the most adequate cluster quality index for data measured on different scales and classified by various clustering methods is well described in literature. Milligan and Cooper (1985) suggest to use Caliński and Harabasz, Hubert and Levine, Baker and Hubert indexes, and also the Silhouette index and the Krzanowski and Lai index are quite commonly used.

The situation differs in case of symbolic data. There are no hints in literature which indexes are most appropriate for that kind of data. This paper describes cluster quality indexes that can be used in this case.

In the first part clustering methods that can be used for symbolic data and methods designed exclusively for this kind of data are described. The second part presents main groups of cluster quality indexes along with examples of indexes from each group (due to the lack of space only the most frequently used indexes are described). The third part describes the classification process for symbolic data. In the next part cluster quality indexes are compared on 100 sets of symbolic data with known structures and for three clustering methods. Furthermore, there is a short summary which of them most accu-

rately represents the structure of the clusters. Finally some conclusions and remarks are given.

2 Clustering methods for symbolic data

Symbolic data, unlike classical data, are more complex than tables of numeric values. Bock and Diday (2000) define five types of symbolic variables:

- single quantitative value,
- categorical value,
- interval,
- multi-valued variable,
- multi-valued variable with weights.

Variables in a symbolic object can also be, regardless of theirs type (Diday (2002)):

- taxonomic representing hierarchical structure,
- hierarchically dependent,
- logically dependent.

A common problem with the usage of symbolic data in classification algorithms is the fact, that for this kind of data, due to their structure, operations of addition, subtraction, multiplication, squaring, calculation of means or calculation of variance are not defined. Thus, methods based on data matrices cannot be used. Only methods based on distance matrices are applicable. Among them the most popular ones are:
Hierarchical agglomerative clustering methods (Gordon (1999, p. 79)):

- Ward,
- single linkage,
- complete linkage,
- average linkage,
- McQuitty (1966),
- centroid,
- median.

Optimization methods:

- Partitioning around medoids, also called k-medoids method (Kaufman and Rousseeuw (1990)).

Algorithms developed for symbolic data (Chavent et al. (2003), Verde (2004)):

- divisive clustering of symbolic objects (DIV),
- clustering of symbolic objects based on distance tables (DCLUST),
- dynamic clustering of symbolic objects (SCLUST),
- hierarchical and pyramidal clustering of symbolic objects (HiPYR).