ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Andrzej Dudek**

INTERNAL CLUSTER QUALITY INDEXES FOR CLASSIFICATION OF SYMBOLIC DATA

Abstract

This paper describes main classification methods used for symbolic data (e.g. data in form of: single quantitative value, categorical value, interval, multivalued variable, multivalued variable with weights) presents difficulties of measuring clustering quality for symbolic data (such as lack of "traditional" data matrix), presents which of known indexes like Silhouette index, Caliński and Harabasz index, Baker and Hubert index, Huberta and Levine index, Ratkovski index, Ball index, Hartigan index, Krzanowski and Lai index, Scott index, Marriot index, Rubin index, Friedman index may be used for validation of such type of data and what indexes are specific only for symbolic data. Simulation results are used to propose most adequate indexes for each classification algorithm.

Key words: Classification, clustering, cluster quality indexes, symbolic data.

1. Introduction

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

Problem of choosing most adequate cluster quality index for data measured on different scales and classified by various clustering methods is well-described in literature. Milligan suggest to use Caliński and Harabasz, Hubert and Levine, Baker and Hubert indexes and also Silhuette index and Krzanowski and Lai indexes are quite commonly used.

* Ph.D., Chair of Econometrics and Informatics, University of Economics, Wrocław.

Situation differs in case of symbolic data (data that can represent numbers, intervals, set of values and qualitative data). There are no suggestions in literature which indexes are most appropriate for these data. This paper describes cluster quality indexes that can be used for symbolic data.

First part is an introduction to symbolic data analysis, symbolic objects and symbolic variables are described and dissimilarity measures for symbolic objects are presented.

In second part clustering methods that can be used for symbolic data and methods specific only for this kind of data are described.

Third part presents main groups of cluster quality indexes along with examples of indexes from each group (due to lack of space only most frequently used indexes are described).

Forth part describes classification process of symbolic data and also an analysis is done which of indexes are calculable for symbolic data.

In next part cluster quality indexes are compared on 20 sets of symbolic data with know structures and with three clustering methods are compared and those of them, which most accurate represents the structure of classes are proposed.

Finally some conclusions and remarks are given.

2. Symbolic objects and symbolic variables

Symbolic data, unlike classical data, are more complex than tables of numeric values. While Table 1 presents usual data representation with objects in rows and variables (attributes) in columns with a number in each cell, Table 2 presents symbolic objects with intervals, set and text data.

Table 1

Classical data situation

X	Variable 1	Variable 2	Variable 3	...
1	1	108	11.98	
2	1.3	123	–23.37	
3	0.9	99	14.35	
...	...	...	...	...

Source: own research.

Table 2

Symbolic data table

X	Variable 1	Variable 2	Variable 3	Variable 4
1	(0.9; 0.9)	{106; 108; 110}	11; 98	{Blue; green}
2	(1; 2)	{123; 124; 125}	–23; 37	{light-grey}
3	(0.9; 1.3)	{100; 102; 99; 97}	14; 35	{pale}
...	...	...	...	...

Source: own research.

Bock and Diday (2000) define five types of symbolic variables:

- single quantitative value,
- categorical value,
- interval,
- multivalued variable,
- multivalued variable with weights.

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

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

Because of the structure of symbolic objects, usual measures like Manhatan distance, Euclidean distance, Canbererra distance or Minkowski metrics cannot be used. For symbolic data, other measures are defined.

There are five main types of dissimilarity measures for symbolic objects (Malerba *et al.*, 2000; Chavent *et al.*, 2003):

- Gowda, Krishna and Diday – mutual neighbourhood value, with no taxonomic variables implemented,
- Ichino and Yaguchi – dissimilarity measure based on operators of Cartesian join and Cartesian meet, which extend operators $\cup$ (sum of sets) and $\cap$ (product of sets) onto all data types represented in symbolic object,
- De Carvalho measures – extension of Ichino and Yaguchi measure based on a comparison function (CF), aggregation function (AF) and description potential of an object,
- Hausdorff distance (for symbolic objects containing intervals),
- L1 distance (Bock and Diday 2000, pp. 302–304).

3. Clustering methods for symbolic data

Common problem in using classification algorithms for symbolic data is fact that for this kind of data due to theirs structure operations of adding, subtracting, multiplying, 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 most popular are:

Hierarchical aggregative clustering methods (Gordon, 1999, p. 79):

- Ward hierarchical clustering,
- single link hierarchical clustering,
- complete link hierarchical clustering,
- average link hierarchical clustering,
- Mcquitty (1966) hierarchical clustering,
- centroid hierarchical clustering.

Optimization methods:

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

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

- Divisive clustering for symbolic objects (DIV),
- Clustering for symbolic object based on distance tables (DCLUST),
- Dynamic clustering for symbolic objects (SCLUST),
- Hierarchical & pyramidal clustering for symbolic objects (HiPYR).

Popular methods like *k-means* and related like *Hard Competitive learning*, *Soft Competitive learning*, *Isodata* and others cannot be used for symbolic data.

4. Cluster quality indexes

Over fifty internal cluster quality indexes are described in literature. Most of them can be arranged in three main groups (Weingessel, *et al.*, 2003), for each group few well-known representatives are enumerated:

Indexes based on inertia (Sum of squares):

- Caliński and Harabasz (1974) index (pseudo F-statistics),
- Hartigan index,
- Ratkovski index,
- Ball index,
- Krzanowski and Lai (1985) index.

Indexes based on scatter matrix:

- Scott index,
- Marriot index,
- Friedman index,
- Rubin index.

Indexes based on distance matrices:

- Silhouette (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990),
- Baker and Hubert (Hubert, 1974; Baker and Hubert, 1975),
- Hubert and Levine (1976).

Different, relatively small, group are indexes dedicated only for symbolic data. Those indexes are (Verde, 2004):

- Inertia for symbolic objects,
- Homogenity index.

5. Clustering quality indexes – symbolic objects case

Figure 1 summarize usage of clustering quality index for symbolic objects. For those objects clustering methods based on data matrices cannot be used. If clustering algorithm is based on distance matrix then for validation based on inertia and indexes based on distance matrix can be used. If algorithm designed strictly for symbolic data is used then for validation indexes based on inertia and "symbolic" indexes are most appropriate.

Four paths of classification procedure may be distinguished for symbolic objects:

- Clustering procedure based on dissimilarity matrix, validation with cluster quality index based on inertia;
- Clustering procedure based on dissimilarity matrix, validation with cluster quality index based on dissimilarity/distance matrix;
- "Symbolic" clustering procedure, validation with cluster quality index based on inertia;
- "Symbolic" clustering procedure, validation with cluster quality index designed for symbolic data.

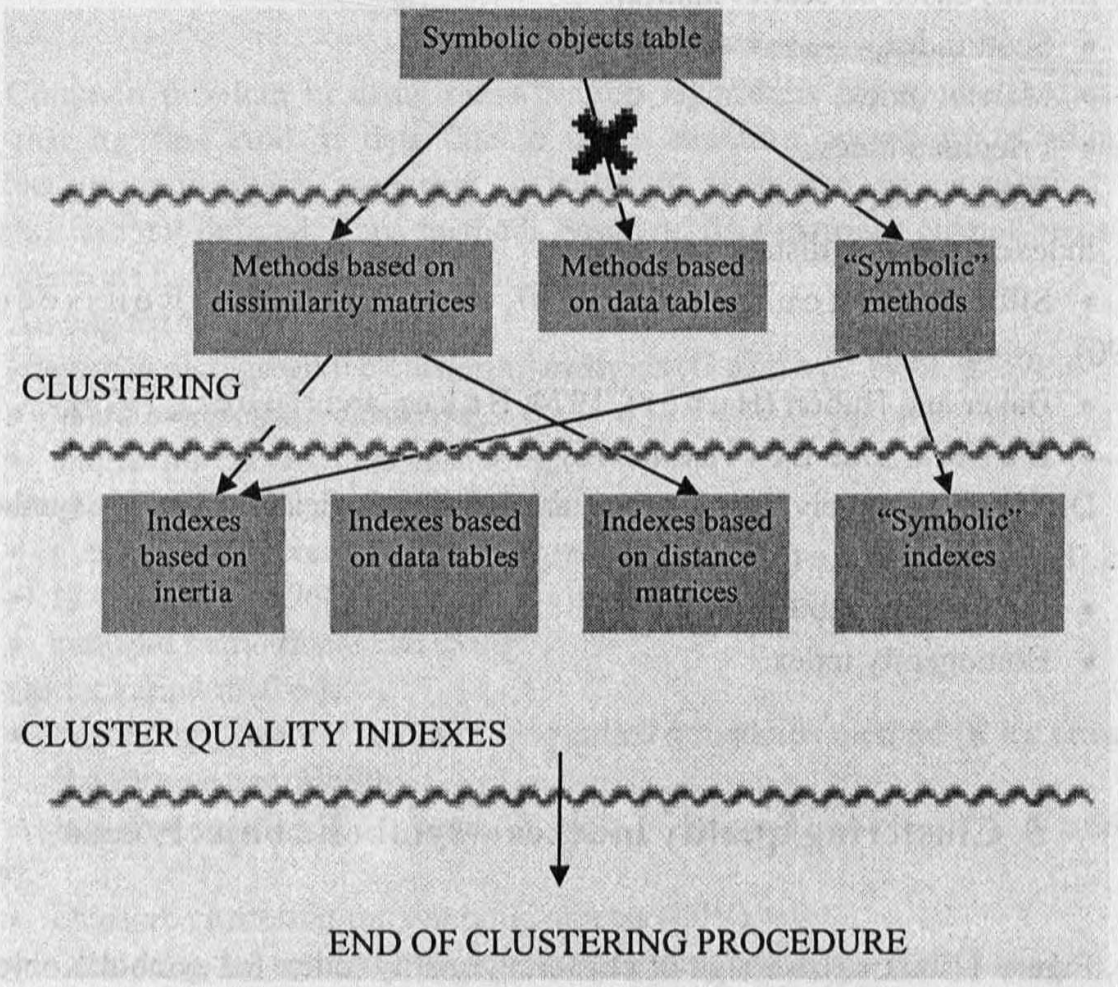

Fig. 1. Clustering method and cluster quality indexes for symbolic data

Source: own research based on Verde 2004, Chavent *et al.*, 2003, Weingessel *et. al.*, 1999.

6. Comparison of clustering quality indexes in symbolic objects case – computational results

Many authors like Milligan and Copper (1985) compared cluster quality indexes and suggested which of them represents real structure of data most adequate. No such comparison has been done for symbolic data yet and an attempt to do so has been made with use of computer program in R environment with symbolic DA library (written in R and C languages by author).

Twenty symbolic data sets with known class structure has been clustered, and compatibility measure for each index has been calculated according to

condition: "If best value of index is achieved for number of cluster corresponding to real structure of data set then compatibility measure is incremented".

Three clustering algorithms has been used: Ward hierarchical clustering method, partitioning around medoids method and dynamical clustering for symbolic objects methods. For each algorithm compatibility measure has been calculated separately.

The following indexes has been compared:

- S – Silhouette index,
- G2 – Baker and Hubert index,
- G3 – Hubert and Levine index,
- F – Caliński and Harabasz index,
- H – Hartigan index,
- SI – inertia for symbolic objects.

Ichino and Yaguchi distance measure was used to calculate distance matrix. Results of the experiment are presented in tables 3–5.

Table 3

Comparison of cluster quality indexes for symbolic data – Ward hierarchical clustering

Index	3 classes Successes (max 4)	4 classes Successes (max 5)	5 classes Successes (max 6)	7 classes Successes (max 5)	Total
S	1	1	0	0	2
G2	3	5	1	3	12
G3	3	3	5	1	12
F	1	1	0	0	2
H	1	5	1	0	7
SI	1	1	2	0	4

Source: own research, calculations made in R environment with use of symbolicDA library.

For ward hierarchical clustering for symbolic objects Hubert and Levine (G3) and Baker Hubert and Hubert (G2) indexes most adequately represent real structure of data. Only Caliński and Harabasz index gives significantly good results and correlation between other indexes values and real class structure is at very low level.

Table 4

Comparison of cluster quality indexes for symbolic data – k-medoids algorithm

Index	3 classes Successes (max 4)	4 classes Successes (max 5)	5 classes Successes (max 6)	7 classes Successes (max 5)	Total
S	3	1	1	0	5
G2	1	3	3	2	9
G3	4	1	5	1	11
F	0	1	0	1	2
H	2	0	0	0	2
SI	4	0	5	0	9

Source: own research, calculations made in R environment with use of symbolicDA library.

For k-medoids algorithm for symbolic objects Hubert and Levine (G2), Baker and Hubert (G3) and symbolic inertia (SI) may be used to validate classification results.

Table 5

Comparison of cluster quality indexes for symbolic data – Dynamical clustering

Index	3 classes Successes (max 4)	4 classes Successes (max 5)	5 classes Successes (max 6)	7 classes Successes (max 5)	Total
S	3	1	1	1	6
G2	1	2	3	3	9
G3	4	1	5	1	11
F	1	0	1	0	2
H	2	0	0	0	2
SI	0	0	1	0	1

Source: own research, calculations made in R environment with use of symbolicDA library.

And again for dynamical clustering for symbolic objects algorithm Hubert and Levine (G3) and Baker and Hubert (G3) indexes most adequately represent real structure of data. Table 6 shows summarized results of experiments. G2 and G3 indexes are significantly better than other indexes.

Table 6

Comparison of cluster quality indexes for symbolic data – aggregated results

Index	3 classes Successes (max 12)	4 classes Successes (max 15)	5 classes Successes (max 18)	7 classes Successes (max 15)	Total
S	7	3	2	1	13
G2	5	10	7	8	30
G3	11	5	15	3	34
F	2	2	1	1	6
H	5	5	1	0	11
SI	5	1	8	0	14

Source: own research, calculations made in R environment with use of symbolicDA library.

7. Final remarks

In this paper several cluster quality indexes were compared for symbolic data. Experiment showed that most adequate for this kind of data are Hubert and Levine and Baker and Hubert indexes.

Note that only one strictly "symbolic" index (e.g. symbolic inertia) has been taken into consideration. Currently new proposals are given (see for example Verde (2004) for symbolic homogenity measure) so this comparison should be repeated when more indexes would be introduced in literature of this subject.

References

Baker F. B., Hubert L. J. (1975), *Measuring the power of hierarchical cluster analysis*, „Journal of the American Statistical Association", **70, 349**, 31–38.

Bock H.-H., Diday E. (eds) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, Springer Verlag, Berlin.

Caliński R. B., Harabasz J. (1974), *A dendrite method for cluster analysis*, "Communications in Statistics", **3**, 1–27.

Chavent M., DeCarvalho F. A. T., Verde R. and Lechevallier Y. (2003), *Trois nouvelle méthodes de classification automatique de données symboliques de type intervalle*, "Revue de Statistique Appliquée" , **LI 4**, 5–29.

Diday E. (2002), *An introduction to symbolic data analysis and the SODAS software*, "J.S.D.A., International E-Journal".

Gordon A. D. (1999), *Classification*, Chapman & Hall/CRC, London.
Hubert L. J. (1974), *Approximate evaluation technique for the single-link and complete-link hierarchical clustering procedures*, "Journal of the American Statistical Association", **69**, 347, 698–704.
Hubert L. J., Levine J. R. (1976), *Evaluating object set partitions: free sort analysis and some generalizations*, "Journal of Verbal Learning and Verbal Behaviour", **15**, 549–570.
Kaufman L., Rousseeuw P. J. (1990), *Finding groups in data: an introduction to cluster analysis*, Wiley, New York.
Krzanowski W. J., Lai Y. T. (1985), *A criterion for determining the number of groups in a data set using sum of squares clustering*, "Biometrics", **44**, 23–34.
Malerba D., Espozito F., Giovalle V., Tamma V. (2001), *Comparing dissimilarity measures for symbolic data analysis*, "New Techniques and Technologies for Statistics" (ETK-NTTS'01), 473–481.
McQuitty L. L. (1966), *Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data*, "Educational and Psychological Measurement", **26**, 825–831.
Milligan G. W., Cooper M. C. (1985), *An examination of procedures for determining the number of clusters in a data set*, "Psychometrika", **2**, 159–179.
Rousseeuw P. J. (1987), *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, "Journal of Computational and Applied Mathematics", **20**, 53–65.
Verde R.(2004), *Clustering methods in symbolic data analysis*, Classification, "Clustering and Data Mining", Berlin-Springer-Verlag, 299–318.
Weingessel A., Dimitriadou A., Dolnicar S. (1999), *An examination of indexes for determining the number of clusters in binary data sets*, available at URL: http://www.wu--wien.ac.at/am/wp99.htm#29.

Andrzej Dudek

Mierniki jakości klasyfikacji dla danych symbolicznych

Artykuł opisuje procedury klasyfikacyjne, które mogą być używane dla danych symbolicznych (tj. dla danych mogących być reprezentowanych w postaci: liczb, danych jakościowych, przedziałów liczbowych, zbioru wartości, zbioru wartości z wagami), przedstawia problemy związane z mierzeniem jakości klasyfikacji dla tych procedur (takie jak brak „klasycznej" macierzy danych) oraz przedstawia, które ze znanych indeksów, takich jak: Silhouette, indeks Calińskiego-Harabasza, indeks Bakera-Huberta, indeks Huberta-Levine, indeks Ratkowskiego, indeks Balla, indeks Hartigana, indeks Krzanowskiego-Lai, indeks Scotta, indeks Marriota, indeks Rubina i indeks Friedmana, mogą być wykorzystane dla tego typu danych oraz jakie są miary jakości podziału specyficzne dla danych symbolicznych. Na podstawie przeprowadzonych symulacji zaproponowane zostały indeksy faktycznie odzwierciedlające strukturę klas dla poszczególnych algorytmów klasyfikacyjnych.