Zbigniew OMIOTEK, Waldemar WÓJCIK

Lublin University of Technology, Institute of Electronics and Information Technology

doi:10.15199/48.2016.11.04

An efficient method for analyzing measurement results on the example of thyroid ultrasound images

Abstract. *The paper presents a method which supports the choice of the clustering procedure and makes it possible to select parameters for most important steps in this process. This method was presented on the example of thyroid ultrasound images belonging to healthy individuals and patients suffering from Hashimoto's thyroiditis. 11 360 variants of clustering procedure were analyzed and optimal parameters for 4 different forms of data set have been chosen.*

Streszczenie. *W pracy zaprezentowano metodę, która wspomaga wybór procedury grupowania obiektów i pozwala określić parametry dla najważniejszych etapów tego procesu. Działanie tej metody pokazano na przykładzie obrazów USG tarczycy należących do osób zdrowych i chorych na chorobę Hashimoto. Metoda pozwoliła przeanalizować 11 360 wariantów procedury grupowania i wybrać optymalne parametry dla czterech różnych postaci zbioru danych. (**Wydajna metoda analizy wyników pomiarów na przykładzie badań USG**).*

Keywords: cluster analysis, clustering procedure, clustering validation, clusterSim.
Słowa kluczowe: analiza skupień, procedura grupowania, ocena grupowania, clusterSim.

Introduction

In many areas of science and technology, measurement results are subsequently used to build a computer recognition system. Depending on the response of such a system for the given input object, we can talk that the system executes a supervised classification, regression or clustering task. We can distinguish 7 steps in a typical grouping procedure: selection of objects and variables, decisions concerning variable normalization formula, selection of a distance measure, selection of clustering method, determining the number of clusters, clustering validation, groups description and profiling. Critical stages are decisions concerning variables normalization formula, selection of a distance measure, selection of clustering method, and determining the number of clusters. These steps are largely arbitrary [1]. Depending on the similarity measure, type of a clustering algorithm and various values of its parameters, we get different splits of a given set of objects. Therefore, such a division of objects into classes is a difficult task and active studies are still carried out on the clustering methods [2-6]. The paper shows the possibility of automated clustering and objective selection of the most important parameters of this process. Our work aims to present the method to accomplish the above task and verifying its usefulness on the example of thyroid ultrasound images.

Research material, tools and methods

In the study, we used series of thyroid ultrasound images belonging to 60 patients. There were 28 healthy patients and 32 patients with a diagnosis of Hashimoto's disease [7, 8]. On this base, we obtained 126 samples belonging to cases identified as sick and 108 samples for healthy cases. A result of the image analysis was a set of 281 image feature descriptors that we reduced using 3 various methods in the next step. We obtained 48 descriptors using the correlation method, 57 descriptors using the *HINoV* method [9], and 3 descriptors by the use of the Hellwig method [10, 11]. During the clustering, the following tools and methods have been used:

- 5 data normalization formulas (classic standardization, Weber standardization, unitarization, zero unitarization, normalization in the interval of [-1; 1];
- 5 distance measures for variables measured in the metric scale (Manhattan, Euclidean, Chebyshev, square Euclidean, generalized distance measure - GDM1);
- simulation method for optimization of the clustering procedure selection (*clusterSim* package was used);
- simulation results were evaluated using 5 indexes of the clustering quality: Caliński and Harabasz, Baker and Hubert, Hubert and Levine, Krzanowski and Lai and Silhouette;
- 9 clustering methods: the nearest neighbor, the furthest neighbor, group average, weighted group average, Ward, centroid, median, *k*-medoids and *k*-means.

The number of variants under consideration of the classification procedure depends on the number of normalization formulas, the number of distance measures and the number of clustering methods. The aforementioned numbers vary depending on a type of the variable measurement scale in a data matrix. Variables used in the study were measured on a quotient and interval scale. For this type of scales and a given index of the clustering quality, the number of variants under consideration of the classification procedure for 7 hierarchical agglomeration methods and *k*-medoids method is equal to 140 (5 standardization formulas, 5 types of the distance measure[1]). In addition, for 2 indexes (Caliński and Harabasz and Krzanowski and Lai) the *k*-means method is used, so the number of variants is further increased by 10 (5 standardization formulas). Because the study included 5 clustering quality indexes, the total number of variants in the analysis of only 1 way of dividing into groups was equal to 710 (5x140 + 2x10). We used such variants for the simulation procedure where the number of groups varied from 2 to 5, therefore the previous number should be multiplied by 4. As a result, the number of variants for 1 type of data set was equal to 2 840. In the analysis 4 types of data sets (full and 3 reduced) were used, therefore the total number of variants under consideration of the classification procedure was 11 360.

Simulation method for optimization of the clustering procedure selection

We used the simulation method to deal with as complex task as the analysis of 11 360 variants of the clustering procedure. For this purpose, the *clusterSim* package written in R has been used. This package consists of a basic *cluster.Sim* function and 16 auxiliary functions. The basic function searches for the optimal clustering procedure

[1] For 3 hierarchical methods (Ward, centroid, median), the squared Euclidean distance as a distance measure is used, because these methods have a geometric interpretation only in this case.

(among various combinations of standardization formulas, distance measures and clustering methods) for a given data type in terms of the chosen clustering quality index. There is a possibility to select of 9 variants of the simulation procedure depending on the variable measurement scale [12]. Individual variants of the clustering process tested in the simulation can be saved in text files (CSV and HTML). Table 1 contains a part of a sample CSV file with simulation results evaluated using the Silhouette index. Similar results were obtained for the other 4 indexes: Caliński and Harabasz, Baker and Hubert, Hubert and Levine and Krzanowski and Lai. It should be added that in the analysis combined training and validation sets were used. Therefore, the evaluation of the clustering quality applies to such a form of data set.

Table 1. A part of a sample CSV file with simulation results

No.	No. of clusters	Normalization formula	Distance measure	Clustering method	Silhouette	Rank
529	2	n5	Squared Euclidean	ward	0.604753907026131	1
557	2	n5	GDM1	pam	0.585327035283877	2
525	2	n5	Squared Euclidean	pam	0.578799494515589	3
559	4	n5	GDM1	pam	0.570046408333300	4
549	2	n5	GDM1	average	0.567865682949157	5
558	3	n5	GDM1	pam	0.559516093684736	6
521	2	n5	Squared Euclidean	mcquitty	0.554006108249404	7
560	5	n5	GDM1	pam	0.550669080867408	8
537	2	n5	Squared Euclidean	median	0.550565688831223	9
38	3	n5	Squared Euclidean	median	0.532990616962180	10

The meaning of the columns in Table 1:

- *No.* – the number of the classification procedure.
- *No. of clusters* – the number of groups.
- *Normalization formula* – a type of the standardization formula (n5 is interpreted by the *cluster.Sim* function as normalization in the interval of [-1, 1]).
- *Distance measure* – a type of the distance measure.
- *Clustering method* – a type of the clustering method.
- *Silhouette* (a name of the index) – a value of the index that specifies the clustering quality. Silhouette index makes it possible to measure the relative compactness and separability of groups and it takes values from the interval of [-1; 1]. Their interpretation in accordance with [13] is as follows: (0.70; 1.00] – strong class structure, (0.50; 0.70] – serious class structure, (0.25; 0.5] – weak class structure, 0.25 and less - no class structure.
- *Rank* – the position of the *i*-th clustering process according with the value of the clustering quality index (1 indicates the best position).

Clustering validation

Simulation results were validated using 3 best global indexes from experiments by Milligan and Cooper [14]: Caliński and Harabasz [15], Baker and Hubert [16] and Hubert and Levine [17], as well as using 2 indexes that are frequently used in the literature for comparative tests: Krzanowski i Lai [18] and Silhouette [13].

To calculate the above-mentioned indexes, the following formulas have been used:

- Caliński and Harabasz index

$$G1(u) = \frac{B_u/(u-1)}{W_u/(n-u)},\ G1(u) \in R_+ \tag{1}$$

- Baker and Hubert index

$$G2(u) = \frac{s(+)-s(-)}{s(+)+s(-)},\ G2(u) \in [-1,1] \tag{2}$$

- Hubert and Levine index

$$G3(u) = \frac{D(u) - r \cdot D_{\min}}{r \cdot D_{\max} - r \cdot D_{\min}},\ G3(u) \in (0,1) \tag{3}$$

- Krzanowski and Lai index

$$KL(u) = \left| \frac{DIFF_u}{DIFF_{u+1}} \right|,\ KL(u) \in R_+ \tag{4}$$

$$DIFF_u = (u-1)^{2/m} W_{u-1} - u^{2/m} W_u$$

- Silhouette index

$$S(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i); b(i)\}},\ S(u) \in [-1,1] \tag{5}$$

where: B_u – matrix of intergroup covariance, W_u – matrix of intragroup covariance, *tr* – matrix trace, $s = 1, ..., u$ – group number, u – the number of groups, $i, k = 1, ..., n$ – object number, n – the number of objects, m – the number of variables, $s(+)$ – the number of pairs of compatible distances, $s(-)$ – the number of pairs incompatible distances, $D(u)$ – the sum of all intragroup distances, r – the number of intragroup distances, D_{min} (D_{max}) – the smallest (largest) intragroup distance, $a(i) = \sum_{k \in \{P_r \setminus i\}} d_{ik} / (n_r - 1)$ – the average distance between the i object and other objects belonging to the P_r group; $b(i) = \min_{s \neq r} \{d_{iP_s}\}$, $d_{iP_s} = \sum_{k \in P_s} d_{ik} / n_s$ – the average distance between the i object and objects belonging to the P_s group.

Indexes $G1(u)$ and $KL(u)$ are based on a data matrix, while $G2(u)$, $G3(u)$ and $S(u)$ on a distance matrix. The maximum value of $G1(u)$, $G3(u)$, $S(u)$ and $KL(u)$ and the minimum value of $G2(u)$ indicate the best division of objects, and they also specify the number of clusters.

Results

The index of Caliński and Harabasz reached the highest value (243) for the data set reduced using the *HINoV* method, and the highest value of the Krzanowski and Lai

index (324) was for the data reduced using the Hellwig method (Fig. 1). In a case of 3 other indexes, the best results were achieved for the *HINoV* method (Fig. 2). The index of Baker and Hubert reached the value of 0.968, Hubert and Levine index was equal to 0.045 (for this index, the smaller value, the better clustering quality), and Silhouette index was equal to 0.741. Table 2 shows optimal variants of the clustering procedure in terms of the individual clustering quality index. For example, for the Caliński and Harabasz index, the largest value (243) was achieved for the number of groups equal to 2, unitarization - as a normalization formula, and *k*-means - as a clustering method.

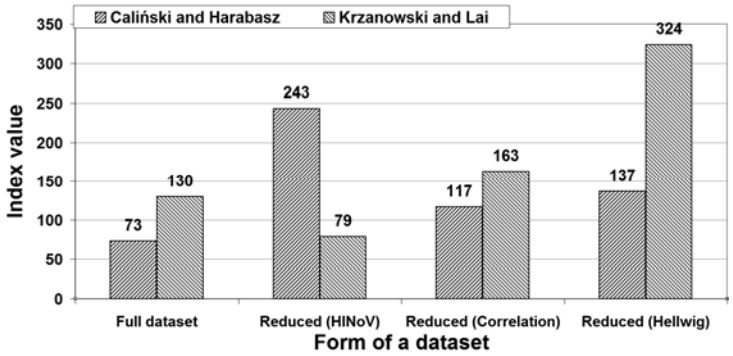

Fig. 1. Calinski & Harabasz and Krzanowski & Lai indexes for various forms of data set

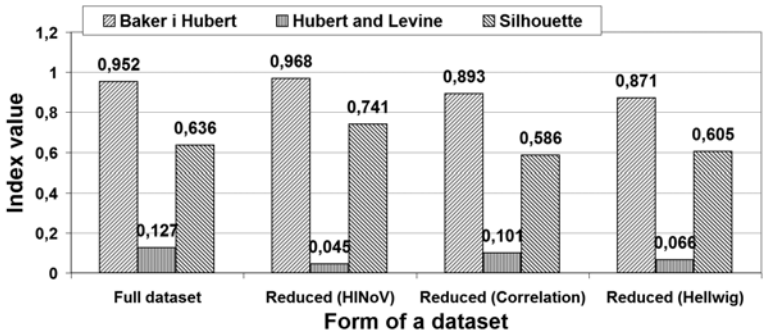

Fig. 2. Baker & Hubert, Hubert & Levine and Silhouette indexes for various forms of data set

Table 2. Optimal variants of the clustering procedure in terms of the individual clustering quality index

Clustering quality index	Clustering procedure parameters	Simulation results	Form of a dataset
Caliński and Harabasz	the number of groups	2	Reduced using *HINoV* method
	variable normalization formula	unitarization	
	distance measure	not applicable	
	clustering method	*k*-means	
Baker and Hubert	the number of groups	2	
	variable normalization formula	normalization in the interval of [-1; 1]	
	distance measure	Chebyshev	
	clustering method	the nearest neighbor	
Hubert and Levine	the number of groups	5	
	variable normalization formula	normalization in the interval of [-1; 1]	
	distance measure	GDM1	
	clustering method	weighted group average	
Silhouette	the number of groups	2	
	variable normalization formula	normalization in the interval of [-1; 1]	
	distance measure	GDM1	
	clustering method	*k*-medoids	
Krzanowski and Lai	the number of groups	2	Reduced using Hellwig method
	variable normalization formula	unitarization	
	distance measure	Manhattan	
	clustering method	weighted group average	

Conclusions

Simulation method for optimization of the clustering procedure selection that was presented in this paper made it possible to select parameters for the most important stages of this process, i.e. variables normalization formula, distance measure, clustering method, and the number of clusters. 2 840 variants of the clustering process for each of the 4 types of data set have been analyzed (11 360 variants in a total).

Parameters that were found provide the optimum division of tested objects into groups in terms of the used clustering quality indexes. It should be noted that results are objective and they have not been burdened with the arbitrary choice of the person leading the study at critical stages of the process.

***Authors**: prof. dr hab. inż. Waldemar Wójcik, dr inż. Zbigniew Omiotek, Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Nadbystrzycka 38d, 20-618 Lublin, E-mail: waldemar.wojcik@pollub.pl; E-mail: z.omiotek@pollub.pl*

REFERENCES

[1] Milligan G.W., Clustering validation: results and implications for applied analyses, [In:] Arabie P., Hubert L.J., de Soete G. (Eds.), *Clustering and Classification. World Scientific* (1996), 341-375

[2] Burda A., Pancerz K., Clustering and Visualization of Bankruptcy Patterns Using the Self-Organizing Maps, *Barometr Regionalny. Analizy i prognozy*, 3 (2014), 133-138

[3] Wosiak A., Zakrzewska D., On Integrating Clustering and Statistical Analysis for Supporting Cardiovascular Disease Diagnosis, *Proceedings of the Federated Conference on Computer Science and Information Systems*, 5 (2015), 303-310

[4] Peker M., A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM, *Journal of Medical Systems*, 40(5) (2016), 116

[5] Anjana Devi M.V., Sarma D.D., Comparison of Clustering Algorithms with Feature Selection on Breast Cancer Dataset, *Journal of Innovation in Computer Science and Engineering*, 5(1) (2015), 59-63

[6] Wu Y., Duan H., Du S., Multiple fuzzy c-means clustering algorithm in medical diagnosis, *Technology and Health Care*, 23 (2015), 519-527

[7] Omiotek Z., Burda A., Wójcik W., The use of decision tree induction and artificial neural networks for automatic diagnosis of Hashimoto's disease, *Expert Systems with Applications*, 40 (2013), n. 16, 6684-6689

[8] Omiotek Z., Burda A., Wójcik W., Application of selected classification methods for detection of Hashimoto's thyroiditis on the basis of ultrasound images, [In:] Pancerz K., Zaitseva E. (Eds.), *Computational Intelligence, Medicine and Biology. Studies in Computational Intelligence*, 600 (2015), 23-37

[9] Carmone F.J., Kara A., Maxwell S., HINoV: a new method to improve market segment definition by identifying noisy rariables, *Journal of Marketing Research*, 36 (1999), 501-509

[10] Hellwig Z., On the optimal choice of predictors. Study VI, [In:] Gostkowski Z. (Ed.), *Toward a system of quantitative indicators of components of human resources development*, UNESCO, Paris, 1968

[11] Omiotek Z., Wójcik W., Zastosowanie metody Hellwiga do redukcji wymiaru przestrzeni cech obrazów USG tarczycy, *Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska*, 3 (2014), 14-17

[12] Walesiak M., Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych - oprogramowanie komputerowe i wyniki badań, *Prace Naukowe AE we Wrocławiu*, 1126 (2006), 120-129

[13] Kaufman L., Rousseeuw P.J., Finding Groups in Data: an Introduction to Cluster Analysis, *Wiley*, 1990

[14] Milligan G.W., Cooper M.C., An Examination of Procedures of Determining the Number of Cluster in a Data Set, *Psychometrika*, 50 (1985), n. 2, 159-179

[15] Caliński R.B., Harabasz J., A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3 (1974), 1-27

[16] Hubert L., Approximate Evaluation Technique for the Single-Link and Complete-Link Hierarchical Clustering Procedures, *Journal of the American Statistical Association*, 69 (1974), n. 347, 698-704

[17] Hubert L.J., Levine J.R., Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations, *Journal of Verbal Learning and Verbal Behaviour*, 15 (1976), 549-570

[18] Krzanowski W.J., Lai Y.T., A Criterion for Determining the Number of Groups in A Data Set Using Sum of Squares Clustering, *Biometrics*, 44 (1985), 23-34