7th International Quality Conference
May 24th 2013
Center for Quality, Faculty of Engineering, University of Kragujevac

Sasa Bogicevic[1)]
Jasmina Vesic Vasovic[1)]
Miroslav Radojicic[1)]
Zoran Necic[1)]

1) University of Kragujevac, Technical faculty, Cacak, Serbia
sasa.bogicevic@hotmail.rs
jasmina.vesic@gmail.com
miroslav.radojicic@yahoo.com
zornes2002@yahoo.com

APPLICATION OF CLUSTER ANALYSIS IN FUNCTION OF IMPROVING DECISION MAKING PROCESS

Abstract: *This paper presents some considerations on the possibilities of improving decision-making by cluster analysis. The paper presents a concrete example of the application of this methodology with appropriate software support. The paper indicates that the discussed methodology have the significant ability for improvement of the quality of decision-making.*

Keywords: *Decision Making, Cluster Analysis*

1. INTRODUCTION

The objective of cluster analysis is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups [1].

There are many reasons for the use of cluster analysis. Used when we analyze the characteristics of the products, selection of employees, in market segmentation etc.

There are a number of clustering methods. Some of them are Ward's method, centroid method, k-means method, linkage methods etc. In this paper a problem is solved using the centroid method.

The problem will be solved using the software package Statistica 8.

2. BASIC TERMS

2.1 Centroid method

In centroid method, clusters are merged on the basis of the Euclidean distance between the cluster centroids. Clusters having least Euclidean distance between their centroids are merged together. In this method, if two unequal sized groups are merged together, then larger of the two tends to dominate the merged cluster. Since centroid methods compare the means of the two clusters, outliers affect it less then most other cluster methods [2].

When solving tasks using cluster methods, the result may be represented graphically. It is a special diagram called a dendrogram.

2.2 Measures of distance for variables

There is more distance metrics [3], and some of them are listed.

Euclidean distance is the most common use of distance – it computes the root of square differences between coordinates of a pair of objects:

$$D_{XY} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}$$

Manhattan distance or city block distance represents distance between points in a city road grid. It computes the absolute differences between coordinates of a pair of objects:

$$D_{XY} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$$

2.3 Standardization of quantitative measures

To determine the similarities between the characters is sometimes convenient to do equalization (transformation) of quantitative measures, to represent all characters in the same way. One of these transformations is the ranking. Dimension x is converted into ranked measure x'. Follows the formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where $x_{\min}$ and $x_{\max}$ minimum or maximum value of characters in a group of taxonomic units that are watching. Ranked measure is always a value between 0 and 1 (because x - $x_{\min}$ is always less than $x_{\max}$ - $x_{\min}$) [4].

3. PROBLEM STATEMENT

Practical application of this method will be shown by an example. Six alternatives will be grouped based on their similarity. We shall consider the six criteria (f_1, f_2, …, f_6).

Table 1. Alternatives

Criteria		Alternatives					
	Requ-est	**a₁**	**a₂**	**a₃**	**a₄**	**a₅**	**a₆**
f₁	max	78	84	95	70	80	92
f₂	max	18	25	28	35	31	30
f₃	max	220	245	215	234	228	240
f₄	max	22	24	16	18	20	14
f₅	max	44	38	50	35	36	41
f₆	max	10	15	12	14	17	12

To the criteria presented in the same way, we rank:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Criterion f1: Criterion f2:

$$x_{a_1} = \frac{78-70}{95-70} = 0{,}32 \qquad x_{a_1} = \frac{18-18}{35-18} = 0$$

$$x_{a_2} = \frac{84-70}{95-70} = 0{,}56 \qquad x_{a_2} = \frac{25-18}{35-18} = 0{,}41$$

$$x_{a_3} = \frac{95-70}{95-70} = 1 \qquad x_{a_3} = 0{,}59$$

$$x_{a_4} = \frac{70-70}{95-70} = 0 \qquad x_{a_4} = 1$$

$$x_{a_5} = \frac{80-70}{95-70} = 0{,}40 \qquad x_{a_5} = 0{,}76$$

$$x_{a_6} = \frac{92-70}{95-70} = 0{,}88 \qquad x_{a_6} = 0{,}71$$

After the transformation of the data, we form a table and execute the cluster analysis.

Table 2. Ranked data

	f₁	**f₂**	**f₃**	**f₄**	**f₅**	**f₆**
a₁	0,32	0	0,17	0,80	0,60	0
a₂	0,56	0,41	1	1	0,20	0,71
a₃	1	0,59	0	0,20	1	0,29
a₄	0	1	0,63	0,40	0	0,57
a₅	0,40	0,76	0,43	0,60	0,07	1
a₆	0,88	0,71	0,83	0	0,40	0,29

Based on the data from the table, we create a distance matrix. In this case we use the Euclidean distance.

$$d(a_x, a_y) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

$$d(a_1, a_2) = \sqrt{\begin{array}{l}(0{,}32-0{,}56)^2 + (0-0{,}41)^2 + \\ (0{,}17-1)^2 + (0{,}80-1)^2 + \\ (0{,}60-0{,}20)^2 (0-0{,}71)^2\end{array}} = 1{,}27$$

$$d(a_1, a_3) = 1{,}20 \qquad d(a_3, a_4) = 1{,}64$$

$$d(a_1,a_4)=1{,}47$$
$$d(a_1,a_5)=1{,}40$$
$$d(a_1,a_6)=1{,}42$$
$$d(a_2,a_3)=1{,}64$$
$$d(a_2,a_4)=1{,}10$$
$$d(a_2,a_5)=0{,}86$$
$$d(a_2,a_6)=1{,}20$$
$$d(a_3,a_5)=1{,}45$$
$$d(a_3,a_6)=1{,}06$$
$$d(a_4,a_5)=0{,}70$$
$$d(a_4,a_6)=1{,}14$$
$$d(a_5,a_6)=1{,}17$$

Now, we are presenting the procedure to obtain the distance matrix using the software package Statistica 8.

First you need to enter data from the table to the program. The next step is shown in the following figure.

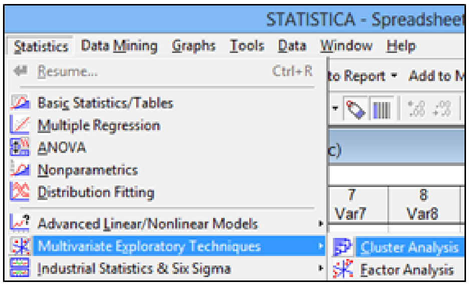
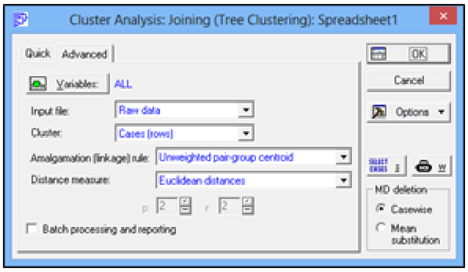


Figure 1. Starting cluster analysis

After that, a new window opens. It selects alternatives, methods, and the distance that we want to use in solving problems. In the "Variables" choose the alternative, in our case, we choose all.


Cluster Analysis: Joining (Tree Clustering): Spreadsheet1

Quick | Advanced

Variables: ALL

Input file: Raw data

Cluster: Cases (rows)

Amalgamation (linkage) rule: Unweighted pair-group centroid

Distance measure: Euclidean distances

p: 2 r: 2

□ Batch processing and reporting

MD deletion: ● Casewise ○ Mean substitution

Figure 2. Selecting method

Confirm to open the window to choose forms of presentation of results, and intermediate results.

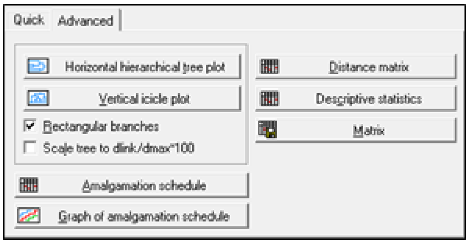
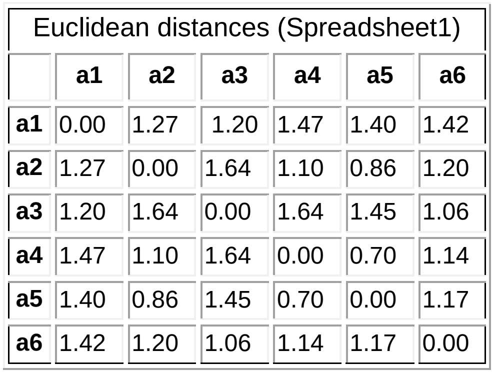


Figure 3. Choice of form results

Clicking on the "Distance matrix" we get the distance matrix, which is analogous to the previous calculations.

Table 3. Distance matrix


Euclidean distances (Spreadsheet1)

	a1	a2	a3	a4	a5	a6
a1	0.00	1.27	1.20	1.47	1.40	1.42
a2	1.27	0.00	1.64	1.10	0.86	1.20
a3	1.20	1.64	0.00	1.64	1.45	1.06
a4	1.47	1.10	1.64	0.00	0.70	1.14
a5	1.40	0.86	1.45	0.70	0.00	1.17
a6	1.42	1.20	1.06	1.14	1.17	0.00

Since the minimum distance between the a4 and a5, they are merged into a new group (cluster).

Centroid of the new group is calculated as the arithmetic mean:

$$c_1=\frac{a_4+a_5}{2}=\frac{(0{,}1{,}0.63{,}0.4{,}0{,}0.57)+(0.40{,}0.76{,}0.43{,}0.60{,}0.07{,}1)}{2}$$
$$=(0.2{,}0.88{,}0.53{,}0.5{,}0.035{,}0.78)$$

Centroid of the new group is calculated as the arithmetic mean:

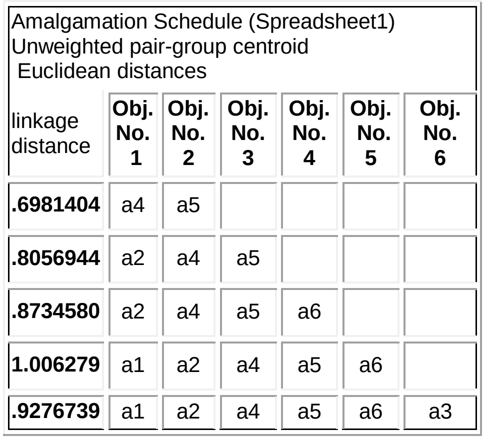
If the problem was solved manually, in the same way would be calculated Euclidean distance. Again to get a new centroid, and this procedure is repeated until the result is one cluster.

Table 4. New data

	f1	f2	f3	f4	f5	f6
a1	0,32	0	0,17	0,80	0,60	0
a2	0,56	0,41	1	1	0,20	0,71
a3	1	0,59	0	0,20	1	0,29
c1	0,2	0,88	0,53	0,50	0,035	0,78
a6	0,88	0,71	0,83	0	0,40	0,29

In the software package Statistica 8 by this method, now we get a table showing the steps in grouping clusters.

Table 5. Steps alternative grouping



Amalgamation Schedule (Spreadsheet1)
Unweighted pair-group centroid
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6
.6981404	a4	a5				
.8056944	a2	a4	a5			
.8734580	a2	a4	a5	a6		
1.006279	a1	a2	a4	a5	a6	
.9276739	a1	a2	a4	a5	a6	a3

In the table we can see that they are grouped taxonomic units a4 and a5 at a distance 0.6981. In the second step are grouped cluster, which is grouped in the previous step, with taxonomic unit a2 at a distance 0.8056 etc.

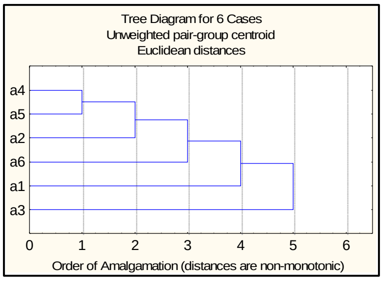


Figure 4. Dendogram

The result of clustering can be displayed with the dendogram, which is most suitable for interpretation.

4. CONCLUSION

In this paper is showed how the alternatives can be grouped, using the method of cluster analysis. This can be very important when choosing alternatives. In this paper, we use the centroid method, but if we use any other method of cluster analysis, the result would be approximately the same.

REFERENCES:

[1] Tryfos, P. (1998). *Methods for Business Analysis and Forecasting: Text &* Cases. New York: Wiley.

[2] Verma, J. P. (2013). *Data analysis in Menagement with SPSS Software*. India: Lakshmibai National University of Physical Education.

[3] Grabust, P. (2011, June). *The choice of metrics for clustering algorithms*. International Scientific and Practical Conference, Rezekne, Latvia.

[4] Tepavčević, A., & Lužanin, Z. (2006). *Mathematical methods in taxonomy*. Novi Sad: Faculty of Sciences.

[5] http://www.statsoft.com

Acknowledgement: Research presented in this paper was supported by Ministry of Education and Science of the Republic of Serbia, Grant III-44010.