COMPARISON OF CLUSTERING RESULTS OF DIFFERENT METHODS WITH FIXED ASSIGNMENT OF OBJECTS TO CLUSTERS

Tomáš Löster

Abstract

Cluster analysis is a popular multivariate method which aims is to create groups of objects. There are many clustering methods that can be used. A very common task of cluster analysis is to find the number of clusters. The aim of this paper is to compare the success of selected coefficients for determining the number of clusters. A total of ten groups of generated files are used for it. There are always three clusters generated, with a total of one hundred files in each of group. In the first group the clusters touch each other. In the second group of files, the clusters are overlap by 10 % of their area. Subsequently, the degree of overlap increases up to level of 90 % and the success of the coefficients is always examined. Based on the performed analyses, it was found, that in case, that clusters are very overlap, the less success rate of coefficients are. From the level overlap of clusters area 20 %, the coefficients are practically inapplicable. The most successful of all used coefficients can be considered the Davies Bouldin coefficient, it´s success rate up to the level of overlap of 20 % is more than 90 %.

Key words: clustering, evaluating of clustering, coefficients, Euclidean distance

JEL Code: C 38, C 40

Introduction

Cluster analysis is a popular multivariate method which aims is to create groups of objects called clusters. It is very often used statistical method, see e.g. Halkidi et al., (2001); Löster (2018, 2019a, b); Řezanková et al., (2013); Bieszk-Stolorz, Dmytrów, (2019); Tatarczak, Boichuk, (2018); Gnat (2019); Objects may be customers, patients, clients, documents, etc. Very often is used to classification of regions. Authors of papers very often used wages to describe regions. The problem of wages and poverty is described e.g. in Bílková, (2012). Other demographic variables, which are very often used in cluster analysis, are described in (2018) or Megyesiová, Rozkošova (2018). In cluster analysis, companies and organizations

are also very often clustered. Ethics is very often analyzed in organizations, such as in Bohinská (2019). In many cases, it is necessary to determine the sample size in order to obtain a sample of the required parameters. Stankovičová, Frankovič (2020) for example solves this problem. Key role in cluster analysis play the similarity characteristics, resp. distances measures. Also in this case, the variable type, which characterizes each object, is critical. In case of quantitative variables the distance measures are used. There are many distance measures between objects. Linkage clustering methods and distance measures a whole series of combinations emerge, the choice is up to the analyst. Various combinations bring different results. In the current literature there are numbers of comparative studies that seek to evaluate various combinations of clustering methods and measure distances in a variety of conditions. However, there is not a clear rule that would strictly determine what combinations use in what situations. Although they are indicated for instance situations in which different distance measures are unsuitable (for example in case of a strong correlation between the input variables), but the actual effect of breaking of this assumption is usually not analyzed. In the same way the advantages and disadvantages of different clustering algorithms are indicated, see Löster, Danko (2019). The aim of the paper is to analyse selected coefficients ant their success rate in finding number of clusters, in different conditions (used methods).

1 Clustering methods and coefficients

In current literature there are many clustering algorithms, which are implemented to many specialized software products. Application of various methods of clustering on same objects described by identical properties can produce different results. Among the methods which were used in this simulation, were included: Nearest neighbour, Farthest neighbour, Average distance, Centroid method and Ward's method. In my simulation was used Euclidean distance measure for clustering. For determining the optimal number of clusters we used four coefficients: Calinski Habarasz (CHF), pseudo-T-squared (PTS) and Davies-Bouldin (DB), Dunn´s coefficients. Detailed descriptions of methods, distances and coefficients can be found e.g. in Řezanková (2009), Gan et al. (2007) or Kogan (2007), Löster (2019b). These coefficients are used very often, see for example Löster, Danko (2019).

2 Groups of files

Ten groups of data files were generated using the random number generator to analyse of success rate of coefficients for finding the number of clusters. There are one hundred files with (generated at the same properties) in all of groups. There are thousand objects divided into three clusters in each cluster. In all cases, the variables are generated from the normal probability distribution. In the first group of files, the clusters touch each other, in the second group the clusters are overlaped 10 %, in the third group the clusters are overlaped 20 %, and consequently the percentage of overlap always increases by 10 % up to the level of 90 %. Selected clustering methods were used to analysis. The number of clusters was determined using selected coefficients and results were compared with the real value of three clusters. On the basis of these analysis, the success rate of the coefficients was determined as a percentage of the number of files for which the correct value was found and the total number of files (100 in every data files).

3 Results

Based on the application of the above clustering methods and the Euclidean distance measure, the following results were obtained for different levels of cluster overlap. Table 1 shows clustering results for group 1 (clusters touching). As we can see from the clustering results, the Davies Bouldin coefficient was the most successful for all used methods. Success rate was 91 % in finding the number of clusters.

Tab. 1: Success rate of coefficients (in%), group of files 1

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	91,00	5,00	13,00	7,00
Farthest neighbour	91,00	1,00	59,00	49,00
Average distance	91,00	6,00	49,00	49,00
Centroid method	91,00	5,00	43,00	43,00
Ward’s method	91,00	5,00	52,00	43,00

Source: our calculations

Table 2 shows clustering results for group 2 (the clusters areas are overlaped 10 %). As we can see from the clustering results, in addition to the Davies Bouldin coefficient, the success rate of other coefficients decreased.

Tab. 2: Success rate of coefficients (in%), group of files 2

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	91,00	0,00	5,00	1,00
Farthest neighbour	91,00	1,00	48,00	39,00
Average distance	91,00	0,00	35,00	35,00
Centroid method	91,00	0,00	37,00	28,00
Ward's method	91,00	2,00	42,00	48,00

Source: our calculations

Table 3 shows clustering results for group 3 (the clusters are overlaped 20 %). As we can see from the clustering results, the success of all coefficients has decreased and is very low.

Tab. 3: Success rate of coefficients (in%), group of files 3

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	12,00	0,00	10,00	1,00
Farthest neighbour	10,00	0,00	27,00	25,00
Average distance	8,00	0,00	10,00	26,00
Centroid method	15,00	0,00	1,00	20,00
Ward's method	8,00	1,00	17,00	25,00

Source: our calculations

Tab. 4: Success rate of coefficients (in%), group of files 4

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	9,00	1,00	7,00	1,00
Farthest neighbour	5,00	1,00	27,00	34,00
Average distance	8,00	0,00	11,00	25,00
Centroid method	12,00	2,00	10,00	19,00
Ward's method	7,00	1,00	7,00	19,00

Source: our calculations

Tab. 5: Success rate of coefficients (in%), group of files 5

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	21,00	1,00	10,00	2,00
Farthest neighbour	11,00	1,00	25,00	23,00
Average distance	7,00	0,00	17,00	22,00
Centroid method	11,00	1,00	12,00	28,00
Ward's method	9,00	2,00	17,00	20,00

Source: our calculations

Tables 4 - 10 show the success of the coefficients in their use in selected clustering methods. There is a gradual increase in the percentage of cluster overlap by 10%. It is obvious that the coefficients at these degrees of overlap are practically unusable for determining the number of clusters.

Tab. 6: Success rate of coefficients (in%), group of files 6

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	21,00	1,00	10,00	2,00
Farthest neighbour	11,00	1,00	25,00	23,00
Average distance	7,00	0,00	17,00	22,00
Centroid method	11,00	1,00	12,00	28,00
Ward's method	9,00	2,00	17,00	20,00

Source: our calculations

Tab. 7: Success rate of coefficients (in%), group of files 7

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	9,00	0,00	10,00	1,00
Farthest neighbour	4,00	0,00	15,00	29,00
Average distance	5,00	0,00	6,00	15,00
Centroid method	12,00	1,00	4,00	15,00
Ward's method	4,00	1,00	9,00	17,00

Source: our calculations

Tab. 8: Success rate of coefficients (in%), group of files 8

Metoda/koeficient	DB	Dunn	CHF	PTS
Nejbližšího souseda	13,00	1,00	10,00	0,00
Nejvzdálenějšího souseda	1,00	0,00	8,00	8,00
Průměrná vazba	0,00	0,00	4,00	17,00
Centroidní metoda	13,00	0,00	4,00	22,00
Wardova metoda	2,00	0,00	1,00	7,00

Source: our calculations

Tab. 9: Success rate of coefficients (in%), group of files 9

Metoda/koeficient	DB	Dunn	CHF	PTS
Nejbližšího souseda	19,00	0,00	7,00	1,00
Nejvzdálenějšího souseda	0,00	0,00	7,00	12,00
Průměrná vazba	0,00	0,00	1,00	8,00
Centroidní metoda	11,00	1,00	3,00	11,00
Wardova metoda	1,00	0,00	0,00	12,00

Source: our calculations

Tab. 10: Success rate of coefficients (in%), group of files 10

Method/coefficient	DB	Dunn	CHF	PTS
Nearest neighbour	12,00	1,00	4,00	0,00
Farthest neighbour	4,00	0,00	14,00	19,00
Average distance	1,00	0,00	4,00	15,00
Centroid method	6,00	0,00	2,00	17,00
Ward's method	2,00	0,00	6,00	15,00

Source: our calculations

Conclusion

The aim of this paper was to compare the success rate of selected coefficients for determining the number of clusters in cluster analysis. Four coefficients were used to analyse - Calinski Habarasz (CHF), pseudo-T-squared (PTS) and Davies-Bouldin (DB), Dunn´s coefficients. Five different clustering methods (Nearest neighbour, Farthest neighbour, Average distance, Centroid method and Ward's method) were used for clustering. In all case was used Euclidean distance measure. One thousand of generated files (from the normal probability distribution) divided into ten groups of files were used for the analysis. In the first group the formed clusters touched each other, in the second group the clusters overlapped 10 %, etc.

Based on the performed analyses, it was found, that in case, that clusters are very overlap, the less success rate of coefficients are. From the level overlap of clusters area 20 %, the coefficients are practically inapplicable. The most successful of all used coefficients can be considered the Davies Bouldin coefficient, it´s success rate up to the level of overlap of 20 % is more than 90 %.

Acknowledgment

This paper was supported by long term institutional support of research activities IP400040 by Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

Bieszk-Stolorz, B., & Dmytrów, K. (2019). Spatial diversity of effectiveness of forms of professional activisation in Poland in years 2008–2014 by poviats. Oeconomia Copernicana, 10(1), 113–130. doi: 10.24136/oc.2019.006

Bílková, D. (2012). Development of wage distribution of the Czech Republic in recent years by highest education attainment and forecasts for 2011 and 2012. In Löster T., Pavelka T. (Eds.), 6th International Days of Statistics and Economics (pp. 162-182). ISBN 978-80-86175-86-7.

Bohinská A. (2019). Compliance program and ethics program: Does an organization need both? In: Journal of Human Resource Management, vol. 22, No. 2, p. 1-9.

Gan, G., Ma, Ch., Wu, J. (2007). *Data Clustering Theory, Algorithms, and Applications*, ASA, Philadelphia.

Gnat, S. (2019). Spatial weight matrix impact on real estate hierarchical clustering in the process of mass valuation. Oeconomia Copernicana, 10(1), 131–151. doi: 10.24136/oc. 2019.007

Halkidi, M., Vazirgiannis, M. (2001). *Clustering validity assessment: Finding the optimal partitioning of a data set*, Proceedings of the IEEE international conference on data mining, pp. 187-194.

Kogan, J. (2007). *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, New York.

Löster, T. (2018). Analysis of Success Rate of the CHF Coefficient in Different Conditions. In: International Days of Statistics and Economics (MSED 2018). Praha, Slaný: Melandrium, Libuše Macáková, pp.1091–1100. https://msed.vse.cz/msed_2018/article/279-Loster-Tomas-paper.pdf.

Löster, T. (2019a). Simulation of the Behavior of Coefficients for Determining the Number of Clusters in Cluster Analysis. In: Aplimat 2019. Bratislava. Publishing house SPEKTRUM STU, pp. 742–749. ISBN 978-80-227-4884-1.

Löster, T. (2019b). *Analýza úspěšnosti vybraných koeficientů pro stanovení počtu shluků*. 1. edition. Slaný: Melandrium, 257 p. ISBN 978-80-87990-17-9.

Löster, T., Danko, J. (2019) Comparison of Success Rate of Selected Coefficients for Determining the optimal Number of Clusters. In: International Days of Statistics and Economics (MSED2019) Praha, Slaný: Libuše Macáková, Melandrium, 2019, pp. 970–977. https://msed.vse.cz/msed_2019/article/231-Loster-Tomas-paper.pdf.

Megyesiová, S., Lieskovská, V. (2018). Analysis of the Sustainable Development Indicators in the OECD Countries. In Sustainability. - Basel: MDPI. ISSN 2071-1050, 2018, vol. 10, no. 12, pp. 1-22

Řezanková, H., Húsek, D., Snášel, V. (2009). *Cluster analysis dat*, 2. edition, Professional Publishing, Praha.

Řezanková, H., Löster, T. (2013). Shlukova analyza domacnosti charakterizovanych kategorialnimi ukazateli. *E+M. Ekonomie a Management*, *16*(3), 139-147. ISSN: 1212-3609.

Stankovičová, I., Frankovič, B. (2020). Určenie veľkosti vzorky pre aplikovaný výskum. In: Forum Statisticum Slovacum, vol. 16, no. 1, p. 57–70.

Tatarczak, A., & Boichuk, O. (2018). The multivariate techniques in evaluation of unemployment analysis of Polish regions. Oeconomia Copernicana, 9(3), 361–380. doi: 10.24136/oc.2018.018

Contact

Ing. Tomáš Löster, Ph.D.
Prague University of Economics and Business
Dept. of Statistics and Probability
W. Churchill sq. 4,
130 67 Prague 3, Czech Republic
tomas.loster@vse.cz