COMPARISON OF SUCCESS RATE OF SELECTED COEFFICIENTS FOR DETERMINING THE OPTIMAL NUMBER OF CLUSTERS

Tomáš Löster – Jakub Danko

Abstract

The aim of this paper is to compare the success rate of selected coefficients for determining the number of clusters in cluster analysis. Three coefficients were selected (CHF, PTS and Davies-Bouldin coefficients). Three different clustering methods were used (Farthest neighbour, centroid and average distance method), two distances were used for clustering (Euclidean and Mahalanobis). Various combinations of clustering methods and distances were created and the given coefficients were applied to determine the number of clusters. We used three hundred of generated files, which were created using a random number generator. In the first group the formed clusters touched each other, in the second group the clusters overlapped 10%, in the third group the clusters overlapped 20%.

Based on the performed analyses, it can be stated that generally most successful coefficient in determining the number of clusters is the CHF coefficient. Its use is usually best when is combined with the farthest neighbour method. When comparing the distance methods used, generally better results are obtained using the Mahalanobis distance. Furthermore, it can be said that with increasing degree of clusters overlap, the success rate of coefficients decreases.

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

JEL Code: C 38, C 40

Introduction

Cluster analysis is multivariate method which main aim is classification of the objects into groups called clusters. It is very often used statistical method, see e.g. Halkidi et al., (2001); Löster at al., (2010); Löster (2018, 2019); Řezanková et al., (2013); Bieszk-Stolorz, Dmytrów, (2019); Tatarczak, Boichuk, (2018); Szymańska, (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) or Marek, (2013). Other demographic variables, which are very often used in cluster analysis, are described in Megyesiová, Lieskovská, (2018) or Megyesiová, Rozkošova (2018). 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. The aim of the paper is to analyse CHF coefficient, which is used for finding number of clusters, in different conditions.

1 Clustering methods

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: the farthest neighbour method, average distance method and centroid method. In our simulation we used two distances - Euclidean and Mahalanobis distance. For determining the optimal number of clusters we used three coefficients: CHF, PTS and Davies-Bouldin (DB) coefficients. Detailed descriptions of methods, distances and coefficients can be found e.g. in Řezanková (2009), Gan et al. (2007) or Kogan (2007).

2 Groups of files

Three groups of data files were generated using the random number generator to analyse of coefficients coefficient. There are one hundred files with the same properties in all groups.

Three clusters are generated in each file. There are thousand objects 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 %. The above clustering methods and both distance measures were applied to these 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 (one hundred).

Table 1 shows clustering results when using the Euclidean Distance Measure for Group 1 (clusters touching). Table 1 shows that the highest success rate was achieved using the furthest neighbor method and using the CHF coefficient. The success rate was 59%.

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

Methods	DB	CHF	PTS
Farthest neighbour	31,00	59,00	49,00
Centroid method	29,00	49,00	49,00
Average distance	36,00	43,00	43,00

Source: our calculations

Table 2 shows clustering results when using the Mahalanobis distance measure for Group 2 (10 % overlaped area). It shows that there was a decline in success compared to when the clusters touched.

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

Methods	DB	CHF	PTS
Farthest neighbour	22,00	48,00	39,00
Centroid method	25,00	35,00	35,00
Average distance	26,00	37,00	28,00

Source: our calculations

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

Methods	DB	CHF	PTS
Farthest neighbour	10,00	27,00	25,00
Centroid method	8,00	10,00	26,00
Average distance	15,00	1,00	20,00

Source: our calculations

Table 3 shows the success rates of coefficients for the group of files No. 3 (clusters overlap 20 % of areas). Obviously, at this degree of overlap, the success rate again decreased over the degree of overlap of 10%.

Table 4 shows clustering results when using Mahalanobis distance for Group 1 (clusters touching). Obviously, the most successful was the DB coefficient, whose success rate using the Farthest Neighbor method was 78 %.

Tab. 4: Success rate of coefficients (in%), group of files 1 (Mahalanobis distance)

Methods	DB	CHF	PTS
Farthest neighbour	78,00	66,00	71,00
Centroid method	41,00	42,00	44,00
Average distance	55,00	60,00	54,00

Source: our calculations

Table 5 shows clustering results when using Mahalanobis distance for Group 2 (10 % overlap of areas). The table shows that the highest success was again achieved using the furthest neighbor method. The success rate of the CHF coefficient was 59 %.

Tab. 5: Success rate of coefficients (in%), group of files 2 (Mahalanobis distance)

Methods	DB	CHF	PTS
Farthest neighbour	52,00	59,00	58,00
Centroid method	33,00	27,00	47,00
Average distance	38,00	50,00	52,00

Source: our calculations

Table 6 shows clustering results when using Mahalanobis distance for Group 3 (20 % overlaped areas of clusters). Again, the table shows that the greatest success was achieved using the furthest neighbor method. The success rate of the CHF coefficient was 50 %.

Tab. 6: Success rate of coefficients (in%), group of files 3 (Mahalanobis distance)

Methods	DB	CHF	PTS
Farthest neighbour	25,00	50,00	43,00
Centroid method	16,00	12,00	36,00
Average distance	26,00	24,00	32,00

Source: our calculations

Tables 7 – 9 show comparisons of coefficient´s success rates for both distance measures. Table 7 shows that in group 1, in most cases better results were obtained using the Mahalanobis distance. The biggest difference was in using of the DB coefficient. Difference was 47% in using of the furthest neighbour method.

Tab. 7: Comparison of results (in%), Group 1, Euclidean and Mahalanobis distance

Methods	DB	CHF	PTS
Farthest neighbour	-47,00	-7,00	-22,00
Centroid method	-12,00	7,00	5,00
Average distance	-19,00	-17,00	-11,00

Source: our calculations

Table 8 shows results for Group of files No. 2 (10 % overlap).

Tab. 8: Comparison of results (in%), Group 2, Euclidean and Mahalanobis distance

Methods	DB	CHF	PTS
Farthest neighbour	-30,00	-11,00	-19,00
Centroid method	-8,00	8,00	-12,00
Average distance	-12,00	-13,00	-24,00

Source: our calculations

The same results, as in the previous cases, were achieved in Group 3 (20 % overlap). Again, better results were obtained using the Mahalanobis distance. The largest difference was found in the CHF coefficient, where the difference between the farthest neighbor method and the average distance method was 23% in using of the Mahalanobis distance measure.

Tab. 9: Comparison of results (in%), Group 3, Euclidean and Mahalanobis distance

Methods	DB	CHF	PTS
Farthest neighbour	-15,00	-23,00	-18,00
Centroid method	-8,00	-2,00	-10,00
Average distance	-11,00	-23,00	-12,00

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. Three coefficients were selected, CHF, PTS and Davies-Bouldin coefficient. Three different clustering methods (farthest neighbour, centroid method, and average distance method) were selected for clustering. Two distances were used for clustering: Euclidean and Mahalanobis. Various combinations of clustering methods and distances were created and the given coefficients were applied to determine the number of clusters. Three hundred of generated files were used for the analysis. In the first group the formed clusters touched each other, in the second group the clusters overlapped 10 %, in the third group the clusters overlapped 20 %.

Based on the performed analyses, it can be stated that generally the most successful in determining the number of clusters is the CHF coefficient. Its use is usually best when combined with the farthest neighbour method. When comparing the distance methods used, generally better results are obtained using the Mahalanobis distance measure. Furthermore, it can be said that with increasing degree of clusters overlap, the success rate of coefficients decreases.

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.

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. Dostupné z: https://msed.vse.cz/msed_2018/article/279-Loster-Tomas-paper.pdf.

Löster, T. (2019). 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.

Marek, L. (2013). Some Aspects of Average Wage Evolution in the Czech Republic. In: International Days of Statistics and Economics. [online], Slaný: Melandrium, pp. 947–958. ISBN 978-80-86175-87-4. URL: http://msed.vse.cz/files/2013/208-Marek-Lubos-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

Megyesiová, S., Rozkošová, A. (2018) Success of Visegrad Group Countries in the Field of Labour Market. - Registrovaný: Scopus. In Journal of Applied Economic Sciences. - Craiova : Spiru Haret University. ISSN 2393-5162, 2018, vol. 13, no. 2, pp. 369-377.

Meloun, M., Militký, J., Hill, M. (2005): Počítačová analýza vícerozměrných dat v příkladech, Academia, Praha.

Řezanková, H., Húsek, D., Snášel, V. (2009). *Cluster analysis dat,* 2. vydání, 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.

Szymańska, A. (2018). National fiscal frameworks in the post-crisis European Union. Equilibrium. Quarterly Journal of Economics and Economic Policy, 13(4), 623–642. doi: 10.24136/eq.2018.030.

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

Ing. Jakub Danko, PhD.
University of Economics, Prague,
Dept. of Statistics and Probability
W. Churchill sq. 4,
130 67 Prague 3, Czech Republic
danj01@vse.cz