Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

COMPARISON OF RESULTS IN DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

LÖSTER Tomáš (CZ)

Abstract. Cluster analysis is a multivariate statistical method which main aim is to classify objects into groups called clusters. Classification of objects into groups is a natural requirement of various scientific disciplines. There are a lot of methods which can be used to classification of objects into clusters, in the current scientific literature. The main aim of this paper is to compare results of cluster number determination in using of different clustering methods. To comparison were used 19 selected real datasets from *The UCI Machine Learning Repository* (http://archive.ics.uci.edu/ml/datasets.html). On the basis of these analyzes, it can be said, that in case of the high level of variable variability, in situation, that the individual clusters are overlaped each other, it is better to use the Mahalanobis distance measure in process of cluster number determination. The best results were obtained using the CHF coefficient and Davides Bouldin index. Their success rate was, especially when using the centroid method, 54, 55 % in the case of the CHF coefficient (54, 55%) and 59, 09 % in case of the Davies-Bouldin index (59, 09%).

Keywords: clustering, evaluating of clustering, methods, coefficients for determining the optimal number of clusters, real datasets

Mathematics Subject Classification: Primary 62H30; Secondary 91C20.

1 Introduction

The basic issue of many scientific disciplines is the modernly so called taxonomy: to organize objects in groups. They may be represented by customers, patients, cars, documents, etc. To create groups there could be used different statistical methods and procedures. If the object (observation) is included into the existing group, there is used the discrimination analysis, if the object is classified into classes that may not be known in advance, there is used the cluster analysis. This is a multivariate statistical method that aim is to create groups of objects, which are called clusters. The objective of cluster analysis is to find objects within one cluster that are as similar as possible and objects from two different clusters should be as least similar as possible. Cluster analysis is very often used statistical method, see e.g. [3], [4], [8], [11], [12]. Very often it is used to regions classification. Authors very often used wages to describe regions. The problem of wages and poverty is described e.g. in [1], [2]. Other demographic variables, which are very often considered in cluster analysis, are described in [10], etc. In the current scientific literature, there are many cluster analysis methods that are used to classify objects. These methods and procedures can be

categorized according to various criteria see e.g. [3], [12].They can be divided into hierarchical and non-hierarchical. They can also be divided into traditional or new approaches.

2 Clustering methods

Among the best known clustering methods can be included the nearest neighbour, furthest neighbour, centroid method, the average distance and also the Ward's method. These methods are included for example in the SYSTAT software and the researcher has also the option to apply the coefficients to determine the optimal number of clusters in connection with these methods, see below.

Nearest neighbour method
Nearest neighbour is the oldest and simplest method. Its origin dates back in 1957 and is associated with the name of P. H. A. Sneath, see [3]. The principle of this method is based on the idea that there is always looking for a pair of the most similar (closest) objects and those objects are associated. Firstly there are found two objects whose distance is the shortest and there is created the first cluster from them. Subsequently there is looking for other object whose distance is the smallest to this cluster. If we analyse the distances between clusters, we have to find the minimum distance of any object in the cluster in relation to any object in another cluster, see [12]. According to professional literature, this method is the easiest, but unfortunately has the disadvantage, which is also known as pipelining. This means the fact that there can be included two objects into one cluster, which are really the closest, however, they are not the closest to the most of other objects. This is due to the fact that a larger number of other objects between them creates the “chain” (also bridge). Another disadvantage is, that this method produces relatively elongate clusters. Formula for adjusting the distance matrix can be found e.g. in [3].

Furthest neighbour method
The author of the furthest neighbour method is Sörensen, see [3]. This method is based on the opposite principle than nearest neighbour. There are connected that two clusters (objects) into one cluster, which have a minimum distance between the most distant objects. The advantage of this method especially is [12], that the resulting clusters are small, consistent, compact, and well separated clusters, and that there is not arises the problem of chaining of the objects of clusters. The detailed formulas for adjusting the distance matrix can be found e.g. in [3].

Average distance method
The average distance method is sometimes called as a compromise method between Closest and Furthest neighbour, see [12]. Their priority is that the results are not affected by the extreme values. The criterion for the formation of clusters is defined as the average distance of all objects in one cluster to all of the objects in the second cluster. In this case, the criterion on whose basis become to the creation of new clusters, is influenced by the values of all the objects in the cluster. There are connected such two clusters, whose average distance is minimal. The formulas for the formal form of the distance matrix can be found e.g. in [3].

Centroid method
Centroid method is associated with the names of Sokal and Michener and it was published under the title "weighted group method", see [3]. The idea of this method is based on the centres of gravity (centroids). There are linked such two clusters whose centroid distance is minimal. In this case the “centroid” is defined as the average of the values of variables in the cluster. As an important

advantage of this method is reported that the results are not affected by outlying objects. On the other hand, as a disadvantage of this method is reported the "inversion", see [3]. There may also arise the "confusing clusters", which means that the distance between the centroid of one pair of clusters may be smaller than the distance between the centroid of another pair, created in the previous step. The formulas for the formal form of the distance matrix can be found e.g. in [3].

Ward's method

Ward's method is associated with two names - Ward and Wishart. Author Ward designed a method for measuring of similarities / dissimilarities of clusters, the author Wishart designed the calculation of Ward's coefficient. The principle of clustering is different from the above mention methods of clustering, in which the distance of clusters were optimized. Ward's method is a process to minimize the heterogeneity of clusters, it means that the clusters are created by maximizing within-groups homogeneity. With this method are small clusters removed and as the result of the clustering process are the clusters with approximately the same number of objects. As noted [3], this method is most commonly used in practice. The criterion of homogeneity of the clusters is the within-group sum of squared deviations from the average (centroid) of the cluster. When connecting the clusters there is the main issue based on the idea, that in every step of clustering must be the smallest increment of Ward's G criterion, see [3].

More detailed description of these methods can be found, for example in [3]. Methods differ, apart from other facts, in the time of formation and also in the way of clustering. Some of them minimize the distance between objects, which means the distance among the closest and furthermost objects. Ward's method solves the principle of clustering based on minimizing the intra-group variability, see [3]. Clustering methods may be combined with various measures of distances. The most famous include Euclidean distance measure or Mahalanobis distance measure, see for example [3]. The result of these combinations is a large number of options that can be used to classify objects.

Part of the cluster analysis is very often the number of clusters determination in which the objects may be distributed. To determine the number of clusters (groups of objects), there is a number of criteria and procedures. As noted above, various techniques for clustering may involve a different distribution of objects into clusters. The issues of determining the number of clusters, in the case of quantitative variables, are the subject of many scientific papers, including for example [4], [5], [6], [7], [8], [9] etc. However, in the current literature there is no clear rule to determine the use of concrete coefficients in different conditions. Various authors create and modify their coefficients and sometimes the coefficients are compared with the existing ones. Among the selected criteria for determining the number of clusters can be included Davies-Bouldin's (DB) index, Dunn's index, RMSSTD index, CHF index and PTS index. These coefficients are already applied e.g. in SYSTAT software and they are commonly used for determining the number of clusters.

For clustering can be used a lot of software products. Among the best-known and most frequently used can be included e.g. IBM SPSS, SAS, STATISTICA, S-PLUS, SYSTAT, STATGRAPHICS, etc. In these are implemented especially the traditional clustering methods, including possible methods of decomposition. Some of them, such as the SAS system do not allow to user to select the appropriate combination of clustering methods and distance measures. The system offers a combination like that. When determining the number of clusters can be used only selected software products. But only the selected coefficients are there implemented. Another widespread software products, such as STATISTICA or STATGRAPHICS does not contain coefficient for determining

the number of clusters. Table 1 summarizes the evaluation and the location of selected coefficients listed above for determining the number of clusters.

Coefficient	Founded extreme	Software
CHF index	maximum	SAS system LE, SYSTAT
PTS index	maximum	SAS system LE, SYSTAT
RMSSTD	minimum	SYSTAT
Davies-Bouldin (DB)	minimum	SYSTAT
Dunn's index	maximum	SYSTAT

Tab. 1. Overview of selected coefficients in software products.

3 Evaluation of coefficients for determining the number of clusters on real data sets

In this part of the paper there are described and evaluated the results of clustering in the existing socio-economic data files. All datasets come from The UCI Machine Learning Repository database (web address: *http://archive.ics.uci.edu/ml/datasets.html*). For the purposes were selected 19 files. Some files are very large (millions of objects). For this reason, were produced individual subgroups, and those were evaluated separately. In this random selection the objects were removed (after the inclusion in a sample file) from the original database, for the reason that the object could not be analysed twice in different random samples. Random sampling was carried out using the SPSS version 20. There were evaluated 22 data files in total, some of which are conventional and well-known classification files. Individual existing files are different in terms of number of clusters, the number of objects and the number of variables. The files also differ by separation of objects, and so in some cases, clusters are significantly overlapped. When selecting files from the database it was controlled, that the files are intended for classification of objects with a possibility to analyse all variables simultaneously, i.e. that to the selection must not enter the files that contain qualitative variables. To subsequent evaluation there were subjected those files for which the number of clusters (classification of object to the cluster) is known, (because of comparison). It is obvious that the result (the number of clusters), which is written in the database, is not always the only one possible classification of objects, but for the purposes of evaluation is the value from the database used. If, for example, will be selected only some variables, it could lead to other initial classification of objects. In the case that in some datasets occurred missing values for some variables, the objects were discarded from further analyses. In the case of unequal units it was carried out standardization using by z-scores. Analysed files (sorted alphabetically):
Abalone, Banknote Authentication, Blood Transfusion Service Center, Cardiotocography, Connectionist Bench (Vowel Recognition - Deterding Data), Energy Efficiency, Glass, Indian Liver Patient, Ionosphere, Iris, Musk (Version 1) QSAR Biodegradation, Statlog (Vehicle Silhouettes) a+b, Susy (extensive file from which was conducted random selection), Vertebral Column 2c, Vertebral Column 3c, Wall-Following Robot Navigation Data, Wholesale Customers, Wine.

Table 2 shows the number of correctly set clusters (in %) at each of the method using Euclidean distance measure. The table shows that the success rate of most coefficients is very low and does not reach 50%.

Method/coefficient	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	9,09	45,45	40,91	36,36	50,00
Farthest neighbour	22,73	18,18	18,18	59,09	36,36
Centroid method	27,27	36,36	22,73	45,45	50,00
Average distance	22,73	31,82	22,73	45,45	45,45
Ward's method	18,18	31,82	31,82	18,18	40,91

Tab. 2. Number of correctly set clusters (in %) – Euclidean distance measure.

Table 3 shows the number of correctly set clusters (in %) at each of the method using Mahalanobis distance measure. It can be seen from the table 3 that the success rate of most of coefficients is higher than the Euclidean distance measure was used.

Method/coefficient	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	4,55	45,45	50,00	50,00	45,45
Farthest neighbour	22,73	31,82	36,36	36,36	36,36
Centroid method	0,00	54,55	40,91	59,09	36,36
Average distance	9,09	50,00	50,00	59,09	50,00
Ward's method	22,73	40,91	22,73	4,55	59,09

Tab. 3. Number of correctly set clusters (in %) – Mahalanobis distance measure.

3 Conclusion

Cluster analysis is a multivariate statistical method which aim is to classify objects into groups called clusters. The objective of cluster analysis is to find objects within one cluster that are as similar as possible and objects from two different clusters should be as least similar as possible. The output of cluster analysis very often contains the number of clusters into which the objects are classified. This information is not known in advance in most real cases. To determine the optimal number of clusters there are many ways (coefficients) that again can be combined with different methods and different measures of distances.

Based on the results were compiled conclusins to Table 4, which contains differences betwwn average succes rate (in %) for both distances. It is evident that success rate is higher for almost of coefficients in using of Mahalanobis distance measure.

Method/coefficient	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	4,55	0,00	-9,09	-13,64	4,55
Farthest neighbour	0,00	-13,64	-18,18	22,73	0,00
Centroid method	27,27	-18,18	-18,18	-13,64	13,64
Average distance	13,64	-18,18	-27,27	-13,64	-4,55
Ward's method	-4,55	-9,09	9,09	13,64	-18,18

Tab. 4. Difference Average success rate (in %) of Euclidean and Mahalanobis distances.

Graphical comparison of differences between success rate of Euclidean and Mahalanobis distance measure is evident from the figure 1.

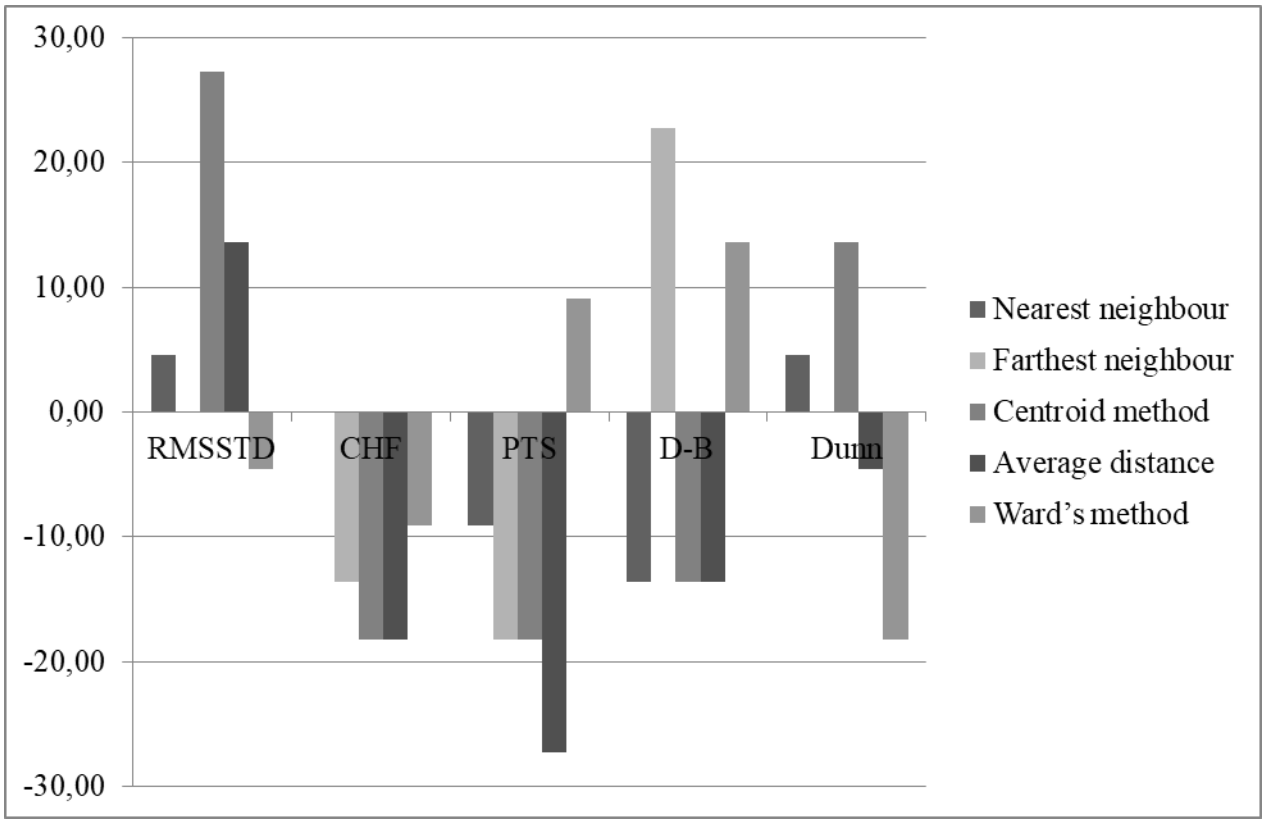

Fig.1. Comparison of differences between success rate of Euclidean and Mahalanobis distance (Eucledian – Mahalanobis).

In conclusion we can say, that the CHF and Davies-Bouldin coefficient are more successful in determining the number of clusters compared with the other coefficients, which are noted e.g. in [6]. The more are the individual clusters overlapped, (i.e. that the separation rate decreases), the less its success rate. Best results are achieved in connection with the average distance method, and Ward's method. The success rate in use of other methods is low. The lowest success of CHF coefficient was achieved in connection with centroid and average distance method. When we compare coefficients in connection with the Mahalanobis distance measure o at the same group of files, it is obvious that the use of Mahalanobis distances extent always leads to a higher success of this coefficient, rather than using Euclidean distance measure. This statement is particularly valid for significantly overlapping clusters.

Acknowledgement

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

References

[1] BÍLKOVA, D. *Modelling of income and wage distribution using the method of l-moments of parameter estimation* . In Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics* (pp. 40-50), 2011, ISBN 978-80-86175-77-5.

[2] BÍLKOVA, D. *Development of wage distribution of the czech republic in recent years by highest education attainment and forecasts for 2011 and 2012*. In Loster Tomas, Pavelka Tomas (Eds.), *6th International Days of Statistics and Economics* (pp. 162-182), 2012, ISBN 978-80-86175-86-7.

[3] GAN, G., MA CH., WU J.: *Data Clustering Theory, Algorithms, and Applications*, ASA, 2007, Philadelphia.

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

[5] LÖSTER, T. *The Evaluation of CHF coefficient in determining the number of clusters using Euclidean distance measure*. In: The 8th International Days of Statistics and Economics., 2014, (pp. 858–869), ISBN 978-80-87990-02-5.

[6] LÖSTER, T. *The Evaluation of CHF coefficient in determining the number of clusters using Mahalanobis distance measure*. In: *14th Conference on Applied Mathematics – Aplimat 2015 [CD]*. Bratislava, 03.02.2015 – 05.02.2015. Bratislava : Slovak University of Technology, 2015, s. 546–554. ISBN 978-80-227-4314-3.

[7] LÖSTER, Tomáš. Evaluation of Coefficients for Determining the Optimal Number of Clusters in Cluster Analysis on Real Data Sets. In: The 9th International Days of Statistics and Economics (MSED 2015) [online]. Praha, 10.09.2015 – 12.09.2015. Slaný : Melandrium, 2015, s. 1014–1023. ISBN 978-80-87990-06-3.

[8] LÖSTER, Tomáš. Evaluation of Coefficients for Determining the Optimal Number of Clusters in Cluster Analysis on Real Data Sets. In: *The 9th International Days of Statistics and Economics (MSED 2015).* [online] Praha, 10.09.2015 – 12.09.2015. Slaný : Melandrium, 2015, s. 1014–1023. ISBN 978-80-87990-06-3.

[9] MAKHALOVÁ, E. 2015. The Fuzzy Clustering Problems and Possible Solutions *The 9th International Days of Statistics and Economics.* Praha, 2015. s. 1052–1061. ISBN 978-80-87990-06-3

[10] MEGYESIOVA, S., & LIESKOVSKA, V. *Recent population change in europe*. In Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics*, (pp. 381-389), 2011, ISBN 978-80-86175-77-5.

[11] MELOUN, M., MILITKÝ, J., HILL, M.: Počítačová analýza vícerozměrných dat v příkladech, Academia, 2005, Praha.

[12] ŘEZANKOVA, H., HÚSEK, D., SNÁŠEL, V.: *Shluková analýza dat,* 2. vydání, Professional Publishing, 2009, Praha.

[13] ŘEZANKOVA, H., & LOSTER, T. Shlukova analyza domacnosti charakterizovanych kategorialnimi ukazateli. *E+M. Ekonomie a Management*, *16*(3), (pp 139-147), 2013, ISSN: 1212-3609.

[14] STANKOVIČOVÁ, I., VOJTKOVÁ, M.: Viacrozmerné štatistické metódy s aplikáciami, Ekonómia, 2007, Bratislava.

Current address

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