ANALYSIS OF SUCCESS RATE OF THE CHF COEFFICIENT IN DIFFERENT CONDITIONS

Tomáš Löster

Abstract

The aim of this paper is to analyse CHF coefficient, which is used for finding the number of clusters in cluster analysis, in different conditions. In current literature there are many methods and many distances measures, which can be used to classification of objects into clusters. In this paper were used different methods (Nearest neighbour, Farthest neighbour, Centroid method, Average distance, Ward's method) in combination with two selected distance measures (Euclidean and Mahalanobis). On the basis of the analyzes it can be stated that if the clusters are overlaped, the ability of the CHF coefficient is lower than in the case of well separated clusters. Success rate was in all cases lower than 60%. The best results are always achieved using the Farthest neighbor method. For the 1st group of files, this it was 48%, respectively, 59 % when was used Euclidean, respectively, Mahalanobis distance measure. For group of files No. 2, the success rate was 46 % resp. 56 %, using Euclidean, respectively, Mahalanobis distance measure. From the above conclusions, it is clear that better results are achieved (in case that 10 % of clusters areas are overlaped) when using the Mahalanobis distance. To compare the probability distribution, better results are obtained in the situation where the variables come from a normal probability distribution.

Key words: clustering, evaluating of clustering, CHF coefficient, Euclidean distance, Mahalanobis distance

JEL Code: C 38, C 40

Introduction

Cluster analysis is multivariate method which objective is to classify the objects into groups called clusters. It is very often used statistical method, see e.g. (Halkidi et al., 2001; Kogan, J., 2007; Řezanková et al., 2013). In practical tasks which are dealing with the classification of objects is crucial for selecting the right multivariate classification methods if they are priory known or unknown the affiliations of the objects into clusters. Objects may be customers,

patients, clients, documents, etc. Authors of papers very often used for example wages to describe regions. The problem of wages and poverty is described e.g. in (Bílková, 2012; Marek, 2013). Other demographic variables, which are very often used in cluster analysis, are described in (Megyesiova, et al. 2011).

In case that objects have known assignment into groups, for classification is used the discriminant analysis. Second situation, i.e. when the classification of the objects is not known in advance is solved by cluster analysis. There are many methods which can be used in current literature.

Key role in cluster analysis play the similarity characteristics, resp. distances measures. Also in this case, the variable type, which characterizes each object, is very important. In case of quantitative variables the distance measures are used. There are many distance measures between objects, which can be used. Various combinations of clustering methods and distance measure provide different results. In the current literature there are a numbers of comparative studies that seek to evaluate various combinations of clustering methods and measure distances in a different conditions. However, there is not a clear rule how to choose best combination in different 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

The aim of cluster analysis is the classification of objects, see (Gan et al., 2007). There are various methods and procedures to do that. These methods and procedures can be categorized according to various criteria see e.g. (Gan et al., 2007; Řezanková et al., 2009; Stankovičová et al., 2007). Mostly they are divided on traditional methods and new approaches in the literature. Traditional methods are well developed and they are applied in many software products.

In current literature there are numbers of clustering algorithms, which are implemented to many specialized software products, see (Meloun et al., 2005). Application of

various methods of clustering on same objects described by identical properties can produce different results. As stated by Gan et al. (2007) and Halkidi (2001) "*It cannot be a priori said which method is the best for a given problem. Usually, the method of the nearest neighbour is the least suitable and method of average distance or Ward's method suits in many cases the best*". Among the methods of hierarchical clustering can be included, for example, the nearest neighbour method, method of the farthest neighbour, method of the average distance, centroid method.

Nearest neighbour method it is the oldest and the simplest method. There are searched two objects, between which the distance is the shortest and they are joined to the cluster. Another cluster is created by linking the third closest object. Distance between two clusters is defined as the shortest distance of any point in cluster in relation to any point in another cluster, see Gan et al. (2007). As one of crucial disadvantage of this methods is stated that occurs so-called *chaining*, when two objects, which are the closest in relation to each other, but not in relation to majority of other objects, are sorted to one cluster.

Farthest neighbour method is based on the opposite principle than the method of the nearest neighbour. The advantage of this method is that it creates small, compact and clearly separated clusters. Contrary to the nearest neighbour method there is no problem with clusters' chaining.

Use of the *method of average distance*, the criterion for emerge of the clusters represents the average distance of all objects in one cluster to all objects located in second cluster. Results of this method are not influenced by extreme values as in the case of method of the nearest and furthers neighbour. Emerge of the cluster is dependent on all objects of clusters.

Centroid method was firstly used by Sokal and Michener under name "weighted group method". This method does not use between-cluster distances of the objects. To new cluster those two clusters are merged, between what is minimal distance of their centroids, while the centroid is understood as an average of the variables in particular clusters. The advantage of this method is that it is not that significantly influenced by remote objects.

Ward's method solves the clustering procedure differently than above stated methods that are optimizing the distances between particular clusters. Method minimizes the heterogeneity of clusters, i.e. clusters are formed using maximization of intragroup homogeneity. As the

measure of homogeneity of clusters is understood intragroup sum of squares of the deviations of values from the average of the clusters and it is called *Ward's criterion*. Criterion for linking the clusters is based on the idea that in each step of clustering there is minimal increment of Ward's criterion. Ward's method has tendency to remove small clusters and create clusters of approximately same size.

Detailed description of these methods can be found, for example, in (Gan, 2007)

Besides the clustering methods themselves and important (key) role is played also by the measures of dissimilarity. Similarity is used as the criterion for the creation of clusters. Measurement of the similarity of objects when they are characterized by quantitative variables is based on the distances of the objects. Transformation of the distance measures to similarity (dissimilarity) measures is done according to simple rules. Very important are the measures of similarities, resp. the distance levels. There are a number of distance levels and in the practice they are combined with various clustering methods, see e.g. (Gan et al., 2007; Řezanková et al. 2009). For measurement of the distances are frequently used:

Euclidean distance represent the length of hypotenuse of a rectangular triangle. Calculation of this measure is based on Pythagoras theorem. *Mahalanobis distance* diminishes the problem while using non-standardized data that can cause differences among clusters due to different measurement units. This measure is usable in the case when all the variables characterizing the objects are mutually correlated.

Detailed descriptions of methods and formulas of particular distance measures can be found e.g. in (Řezanková et. al., 2009) or (Gan, et. al., 2007).

CHF index (also the pseudo *F* index) was designed by the authors Calinski and Habarasz, see (Calinski and Habarasz, 1974). The CHF index is used for finding number of clusters and is defined as the ratio of the average between cluster variability and average within cluster variability. This coefficient represents the *F*-test analogy used in the analysis of variance. It can be used to determine the optimal number of clusters. High values of this coefficient indicate well-separated clusters, i.e., when determining the optimum number of clusters, the maximum value of this index is searched within a predetermined number of clusters.

Detailed descriptions and formulas of this coefficient can be found e.g. in (Calinski et. al., 1974) or (Gan, et. al., 2007).

2 Groups of files

In order to analyze the behavior of the CHF coefficient, a total of two groups of data files were generated using the random number generator. In both groups, there are always hundreds of files with where generated in the same conditions. In each file, there are always three clusters of the same number of thousands of objects. In the first group of files, objects are characterized by two variables that were generated from normal probability distribution. In the second group, there are objects which are characterized by two variables that are uniformed distributed. For both sets of files, the generated clusters overlap. The overlapping area of clusters is 10 % in both groups of files.

For these two groups, the above-mentioned clustering methods were applied, both distance measures and the number of clusters was founded using the CHF coefficient. The results for both groups are compared. Each success rate of CHF for a given combination was determined as a proportion of the number of files that were just 3 clusters founded and the total number of files in that group, which was in both cases one hundred.

Table 1 contains the clustering results when using the Euclidean distance measure in clustering of files in group No. 1.

Tab. 1: Success rate of CHF coefficient (in %), group 1 (Euclidean distance)

Methods	Success
Nearest neighbour	5,00
Farthest neighbour	48,00
Centroid method	35,00
Average distance	37,00
Ward's method	42,00

Source: our calculations

Table 1 shows that the highest success rate was achieved using the Farthes neighbor method.

Table 2 shows the clutering results when was usede the Mahalanobis distance measure at group No. 1.

Tab. 2: Success rate of CHF coefficient (in %), group 1 (Mahalanobis distance)

Methods	Success
Nearest neighbour	6,00
Farthest neighbour	59,00
Centroid method	27,00
Average distance	50,00
Ward's method	39,00

Source: our calculations

As can be seen from table 2, the highest success rate was achieved again using the farthest neighbor method.

Tab. 3: Difference in CHF coefficient success (in %), Group 1 (Euclidean - Mahalanobis distance)

Methods	Success
Nearest neighbour	-1,00
Farthest neighbour	-11,00
Centroid method	8,00
Average distance	-13,00
Ward's method	3,00

Source: our calculations

Table 3 shows the differences in CHF success rates between both distances. These values were determined as the difference between the Euclidean and Mahalanobis distances. Obviously, in most cases, better results have been achieved in using of the Mahalanobis distance. The highest difference was achieved in the average distance method, where the difference was 13 %.

Table 4 shows the resulst of clustering when was the Euclidean distance measure used in clustering of files from group No. 2.

Tab. 4: Success rate of CHF coefficient (in %), group 2 (Euclidean distance)

Methods	Success
Nearest neighbour	7,00
Farthest neighbour	46,00
Centroid method	26,00
Average distance	34,00
Ward's method	28,00

Source: our calculations

As can be seen from the values in Table 4, the highest success rate of the CHF coefficient was again achieved using the most Farthest neighbor method. This success rate is 46 %.

Table 5 shows the clustering results when using the Mahalanobis distance measure in files from group No. 2.

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

Methods	Success
Nearest neighbour	12,00
Farthest neighbour	56,00
Centroid method	26,00
Average distance	44,00
Ward's method	30,00

Source: our calculations

From the values in table 5, it follows that, using the Mahalanobis distance, the highest success rate was achieved again using the most Farthest neighbor method.

Tab. 6: Difference in success rate of CHF coefficient (in %), Group 2 (Euclidean - Mahalanobis distance)

Methods	Success
Nearest neighbour	-5,00
Farthest neighbour	-10,00
Centroid method	0,00
Average distance	-10,00
Ward's method	-2,00

Source: our calculations

Table 6 shows the differences in success rates for both distance measures. The highest difference is achieved with the Farthes neighbor method and the average distance method. It can be stated that better results are achieved in using of the Mahalanobis distance.

Tables 7 and 8 shows the comparison of the success rate of the CHF coefficient for the two groups of sets (different probability distributions). Table 7 shows the results using the Euclidean distance, in Table 8 there are results for using the Mahalanobis distance. Values are always calculated as the success rate of the CHF coefficient for group 1 - success rate for group 2.

Tab. 7: Comparison of results (in%), groups 1 and 2, Euclidean distance

Methods	Success
Nearest neighbour	-2,00
Farthest neighbour	2,00
Centroid method	9,00
Average distance	3,00
Ward's method	14,00

Source: our calculations

Table 7 shows that higher results of success rate were obtained when the variables are generated from a normal probability distribution.

Tab. 8: Comparison of results (in%), groups 1 and 2, Mahalanobis distance

Methods	Success
Nearest neighbour	-6,00
Farthest neighbour	3,00
Centroid method	1,00
Average distance	6,00
Ward's method	9,00

Source: our calculations

Table 8 shows that the higher success rates in using of the Mahalanobis distance were again achieved in situation, that the variables are generated from a normal probability distribution.

Conclusion

The aim of this article was to analyze the behavior of the CHF coefficient on two groups of files. Both groups are artificially generated files, of which the first set of files is generated from a normal probability distribution, the second group is uniformed distributed. A total of five clustering methods were applied to both groups and two distance measure were used. In both groups of files, there are always three clusters that are overlapping. 10 % of clusters area are overlapped in all data sets. In each of clusters there is a thousand objects.

On the basis of the analyses can be stated that in situation of overlapped clusters success rate of the CHF coefficient is lower than in the case of well separated clusters. Success rate in all cases is lower than 60%. The best results are always achieved using the Farthest neighbor method. For the 1st group of files, this is 48 %, respectively 59 % when was used Euklidean, respectively, Mahalanobis distance measure. For group No. 2, the success rate was 46 % resp. 56 % in using Euclidean, respectively, Mahalanobis distance measure. From the above conclusions, it is clear that better results are achieved, in case of overlapped clusters, in used of the Mahalanobis distance measure. To compare the probability distribution, better results are obtained in a situation where the variables come from a normal probability distribution.

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

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.

Calinksi, T., Harabasz., J. (1974). A Dendrite Method for Cluster Analysis. *Comunications in Statistics*, No. 3, s. 1-27.

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

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.

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.

Megyesiova, S., & Lieskovska, V. (2011). Recent population change in Europe. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 381-389). ISBN 978-80-86175-77-5.

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.

Stankovičová, I., Vojtková, M. (2007): Viacrozmerné štatistické metódy s aplikáciami, Ekonómia, Bratislava.

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
losterto@vse.cz