THE EVALUATION OF CHF COEFFICIENT IN DETERMINING THE NUMBER OF CLUSTERS USING EUCLIDEAN DISTANCE MEASURE

Tomáš Löster

Abstract

There are many methods of clustering in current literature and there is possible to use the various measures of distances (resp. similarities), and therefore the resulting distributions of objects into the clusters may be different. There is no strict rule in the literature to determine which method is necessary to use and in which conditions. Also, one part of the cluster analysis is very often to determine the number of clusters. The aim of this paper is to evaluate the CHF coefficient, which is often used to determine the number of clusters. There will be used 20 artificially created files for clustering. Conditions are, that the clusters must be touched or partially overlapped. These generated files are created under the same conditions in order to consider the objective results. Based on analyses it was found, that the CHF coefficient is very successful in determining the number of clusters. In the case that the clusters are touched to each other, its success is 100 % at the generated files. In the case that the clusters are partially overlapped, its success decreases. The highest success in the case of partially overlapped clusters was 70 %. It can be concluded, that the lower rate of separation is, (i.e. the more individual clusters are overlapped), the lower is the success of this coefficient in order to determine the number of clusters.

Key words: clustering, evaluating of clustering, methods, CHF coefficient

JEL Code: C 38, C 40

Introduction

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, 2007; Rezankova et al., 2009). 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. 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, 2007; Rezankova et al. 2009) Very frequently used is the Euclidean distance measure. In the context of this article we will consider only this one distance measure. We will examine which results are achieved when we determine the number of clusters together with the various methods of clustering using the coefficient of CHF. Cluster analysis is very often used statistical method, see e.g. (Halkidi et al., 2001; Loster at al., 2010; Rezankova et al., 2013; Žambochová, 2012). Very often is used to classification of regions. Authors of papers very often used wages to describe regions. The problem of wages and powerty is described e.g. in (Bílková, 2011, 2012; Marek, 2013; Pavelka, 2012; Miskolczi, 2011, Želinský, 2012). Other demographic variables, which are very often used in cluster analysis, are described in (Megyesiova, et al. 2011, 2012).

1 Euclidean distance

Euclidean distance (also known as geometric metric) is the most frequently used measure of distance. It represents the length of the hypotenuse of a right-angled triangle. The calculation of the Euclidean distance measure of ith and jth object is based on the Pythagorean Theorem according to the formula:

$$D_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{t} \left(x_{il} - x_{jl}\right)^2} \,. \qquad (1)$$

Euclidean distance is a standard distance type. When we use the Ward's clustering method, (which will be defined below), there are typically used the squares of this distance. Euclidean distance is not suitable for the case where the individual variables (that characterize the individual characters) are strongly correlated with each other, see (Gan, 2007).

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.

2.1 Nearest neighbour method (single linkage)

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 (Rezankova, et al., 2009). 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 (Meloun, et al., 2005). 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, as noted e.g. (Stankovičová, et al. 2007) is, that this method produces relatively elongate clusters. Formula for adjusting the distance matrix can be found e.g. in (Gan, et al., 2007).

2.2 Furthest neighbour method (complete linkage)

The author of the farthest neighbour method is Sörensen, see (Rezankova, 2007). 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 (see e.g. (Stankovičová, 2007)), 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 (Gan, et al., 2007).

2.3 Average distance method (average linkage, Sokal-Sneath method)

The average distance method is sometimes called as a compromise method between Closest and Farthest neighbour, see (Stankovičová, 2007). 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 (Gan et al., 2007).

2.4 Centroid method (Gower method, the gravity centre method)

Centroid method is associated with the names of Sokal and Michener and it was published under the title "weighted group method", see (Gan et al., 2007). 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 (Meloun et al., 2005). 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 (Gan, et al., 2007).

2.5 Ward's method (Ward - Wishart 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 (Stankovičová, 2007), 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 (Gan, et al., 2007).

2.6 CHF coefficient

To determine the optimal number of clusters we can use the different coefficients, see ... This paper focuses on the evaluation of CHF coefficient (also called pseudo F index), which is very often used in practice. It is based on the decomposition of the total sum of squares and it was designed by the authors Calinski and Habarasz, see (Rezanková, et al., 2009), furthermore, it was studied by the authors Maulik and Bandyopadhyay. CHF index is defined as the ratio of the average between-group and within-group variability, i.e. by the formula

$$I_{\mathrm{CHF}}(k) = \frac{\dfrac{SS_{\mathrm{B}}}{k-1}}{\dfrac{SS_{\mathrm{W}}}{n-k}} = \frac{(n-k)\cdot SS_{\mathrm{B}}}{(k-1)\cdot SS_{\mathrm{W}}}, \tag{2}$$

where

SS_B = sum of squares between clusters (the characteristic of between-group variability),

SS_W = sum of squares within clusters (the characteristic of within-group variability),

SS_T = total sum of squares (the characteristic of total variability).

The individual sums of squares are determined by the following formulas:

$$SS_{\mathrm{W}} = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{t=1}^{m} (x_{it} - \bar{x}_{ht})^2, \tag{3}$$

$$SS_{\mathrm{T}} = \sum_{i=1}^{n} \sum_{t=1}^{m} (x_{it} - \bar{x}_{t})^2, \tag{4}$$

$$SS_{\mathrm{B}} = SS_{\mathrm{T}} - SS_{\mathrm{W}}, \tag{5}$$

where

n is the number of objects,

m is the number of variables which characterize objects,

k is the number of clusters,

$\bar{\mathbf{x}}_h$ is the centroid of h^{th} cluster, and

x_{it} is the value of the t^{th} variable for the i^{th} object.

This coefficient represents the analogy of F-test, which is used in the analysis of variance. It can be used to determine the optimal number of clusters *k**. High values of this coefficient indicates the well separated clusters, it means that during the determine the optimal number of clusters is looking for the maximum value of this index within a predetermined number of clusters

$$I_{\mathrm{CHF}}(k^*) = \max_{2 \le k \le n-1} I_{\mathrm{CHF}}(k). \qquad (6)$$

3 Evaluation of CHF coefficient

For the evaluation of CHF coefficient was necessary to generate a sufficient amount of data files that satisfy the same conditions, in order to objective evaluation of the ability of the coefficient for the determination of optimal number of clusters. The input parameters of the generator are: the number of objects in each of the clusters (equal in all clusters), the number of generated clusters, the number of variables which characterize the objects, variability within the cluster and separation of clusters. (The clusters are well separated, they may be either touched or they may be partially or significantly overlapped). The generator principle is as follows: firstly, in random m-dimensional space (according to number of selected variables) place the centroids of k clusters. Next, on the basis of the separation of clusters (the distance between the centroids of clusters) and on the basis of variability within the clusters, place randomly the selected numbers of objects around the centroids of clusters. For the purpose of examining CHF coefficient there were created two groups of files. Both groups contain 10 files. Typical for them is that the generated clusters are touched or partially overlapped.

3.1 GROUP 1

Within this group of files there were generated five clusters for each set. Each cluster contains 175 objects, each of which is characterized by three quantitative variables. For this group of files is important, that the resulting clusters are touched but not overlapped. Table 1 shows the number of correctly set clusters using CHF coefficient at each of the methods. It is evident that in the case that the resulting clusters are touched but not overlapped, the CHF coefficient is applicable with a high success rate.

Tab. 1: Number of correctly set clusters

Method/Coefficient	CHF	Number of cases
Nearest neighbour	10	10
Farthest neighbour	10	10
Centroid method	10	10
Average distance	10	10
Ward's method	10	10

Source: our calculation

3.2 GROUP 2

Within this group of files there were generated the files, which each contain always three clusters. Each cluster contains 200 objects and each of which is characterized by three quantitative variables. Individual clusters are partially overlap, i.e. that some objects are located in the space of one cluster and should be located in the space of another cluster as well. The resulting number of clusters, which were set on the basis of CHF coefficient for each method are shown in Table 2.

Tab. 2: Number of correctly set clusters

Method/Coefficient	CHF	Number of cases
Nearest neighbour	4	10
Farthest neighbour	7	10
Centroid method	5	10
Average distance	7	10
Ward's method	6	10

Source: our calculation

The table shows that CHF coefficient has a lower percentage of success in determining the correct number of clusters in the case when the clusters are overlapped. For example, when we use Ward's method, this percentage was 6 of 10, respectively in the case of the farthest neighbour method and method of the average distance it was 7 of 10, which is lower success rate compared to the situation, when the clusters were touched.

Conclusion

Based on the results of two groups of files, that were generated under the same conditions in order to consider results as objective, there were compiled the Table 3, which contains a summary of the evaluation of the success of CHF coefficient. In total, there were performed 100 clustering (each file was clustered using five different methods in connection with the square Euclidean distances), and using the CHF coefficient there was determined the number of clusters. For one group of files is characteristic that the resulting clusters may touched each other and the second group of files is characteristic that they may be partly overlapped, which is a frequent situation in practice.

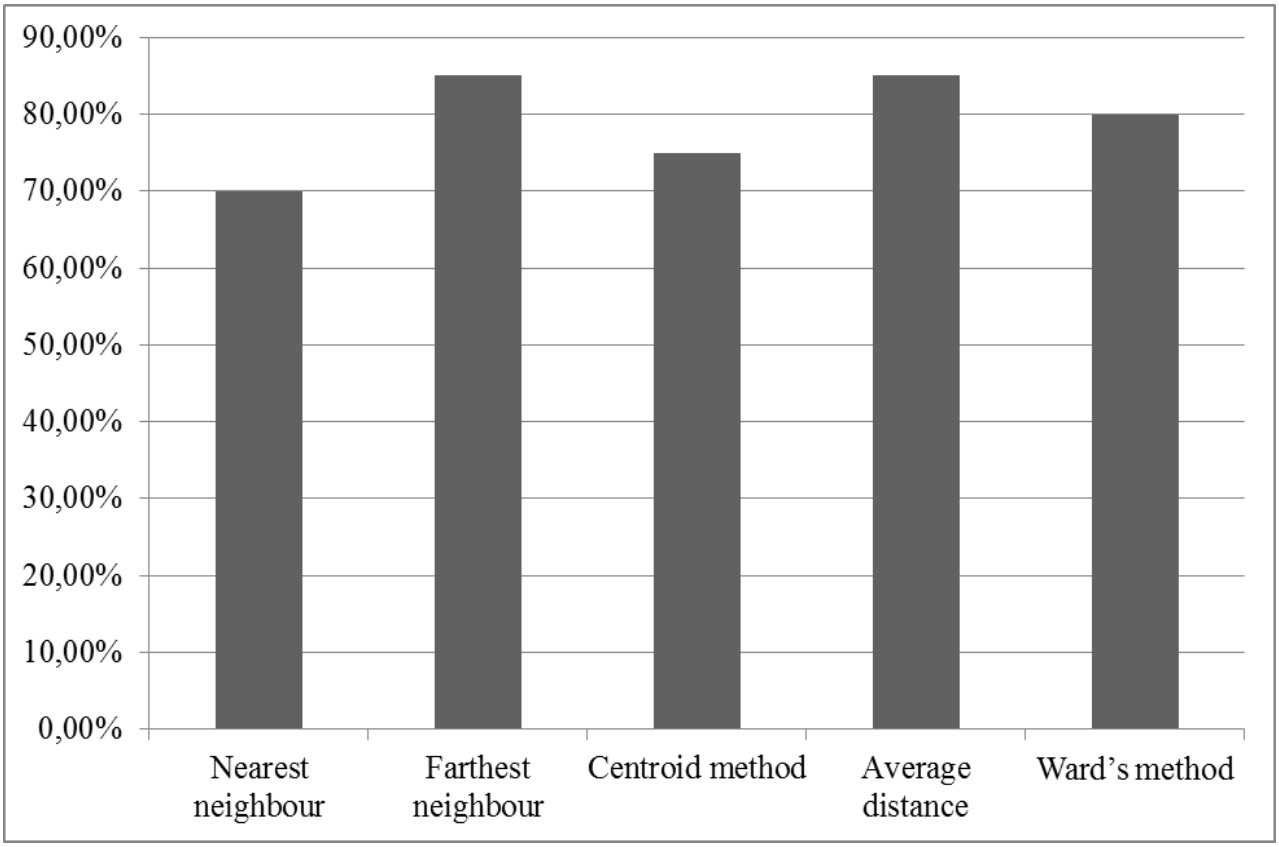
Tab 3: Percentage of correctly set clusters for the whole group

Method/Coefficient	CHF
Nearest neighbour	70,00 %
Farthest neighbour	85,00 %
Centroid method	75,00 %
Average distance	85,00 %
Ward’s method	80,00 %

Source: our calculation

Table 3 shows that the highest success of CHF coefficient have been achieved in connection with the furthest neighbour method, and average distance method. In both cases, the percentage was 85 %. Conversely, the lowest success of CHF coefficient was achieved in connection with the nearest neighbour method. In conclusion we can say, that the CHF coefficient is very successful in determining the number of clusters compared with the other coefficients, which are noted e.g. in (Gan et al., 2007), however, the more are the individual clusters overlapped, (i.e. that the separation rate decreases), thereby also decreasing its success. Graphical display of the success rate of CHF coefficient is also seen from Figure 1.

Fig. 1: The success of the CHF coefficient



Source: our calculation

Acknowledgment

This article was created with the help of the Internal Grant Agency of University of Economics in Prague No. 6/2013 under the title „Evaluation of results of cluster analysis in Economic issues."

References

Bilkova, D. (2011). *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). ISBN 978-80-86175-77-5.

Bilkova, D. (2012). *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). ISBN 978-80-86175-86-7.

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, s. 187-194.

Loster, T., & Langhamrova, J. (2011). *Analysis of long-term unemployment in the czech republic* . In Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics* (pp. 307-316). ISBN 978-80-86175-77-5.

Marek, L. (2013). *Some Aspects of Average Wage Evolution in the Czech Republic*. In: International Days of Statistics and Economics. [online], Slaný: Melandrium, s. 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 Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics* (pp. 381-389). ISBN 978-80-86175-77-5.

Megyesiova, S., & Lieskovska, V. (2012). *Are europeans living longer and healthier lives?*. In Loster Tomas, Pavelka Tomas (Eds.), *6th International Days of Statistics and Economics* (pp. 766-775). ISBN 978-80-86175-86-7.

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

Miskolczi, M., Langhamrova, J., & Fiala, T. (2011). *Unemployment and gdp*. In Loster Tomas, Pavelka Tomas (Eds.), *International Days of Statistics and Economics* (pp. 407-415). ISBN 978-80-86175-77-5.

Pavelka, T. (2012). *The minimum wage in the czech republic – the instrument for motivation to work?*. In Loster Tomas, Pavelka Tomas (Eds.), *6th International Days of Statistics and Economics* (pp. 903-911). ISBN 978-80-86175-86-7.

Rezankova, H., Húsek, D., Snášel, V. (2009): *Shluková analýza dat,* 2. vydání, Professional Publishing, Praha.

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

Simpach, O. (2012). *Statistical view of the curent situation of beekeeping in the czech republic*. In Loster Tomas, Pavelka Tomas (Eds.), *6th International Days of Statistics and Economics* (pp. 1054-1062). ISBN 978-80-86175-86-7.

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

Žambochová, M. (2012): *Classification in terms of students' preferences For information sources, Efficiency and responsibility in education.* 9th International Conference on Efficiency and Responsibility in Education, Praha, s. 612-620, ISBN 978-80-213-2289-9.

Zelinsky, T., & Stankovicova, I. (2012). *Spatial aspects of poverty in slovakia*. In Loster Tomas, Pavelka Tomas (Eds.), *6th International Days of Statistics and Economics* (pp. 1228-1235). ISBN 978-80-86175-86-7.

Contact

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