EVALUATION OF COEFFICIENTS FOR DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTER ANALYSIS ON REAL DATA SETS

Tomáš Löster

Abstract

Cluster analysis is a multivariate method, which aim is classification of objects. The main aim is that objects within groups (clusters) are the most similar, and objects, which are in two different clusters, are the least similar. In the current literature, there are many methods and many measures of distances that can be combined. Various combinations of methods and distances give different results. In the current literature, there is not rule for choosing of these combinations. At the same time, it is necessary to determine the optimal number of clusters into which the clusters will be classified. Even in this case it is not clearly addressed in conjunction with specific method and distance should be index used. This article aims to evaluate the selected coefficients for determining the number of clusters in combinations with different methods and with different distance measures. Based on the analysis of 32 existing data files from the database *The UCI Machine Learning Repository* been found that the success of different coefficients for determining the number of clusters is not only different for different clustering methods, but also in combination with different distance measures. For example, CHF coefficient is preferable to use in combination with the Mahalanobis distance, where the success rate is higher compared to the Euclidean distance. For example, when using the method average distance, success rate of this coefficient is higher by 21, 88%. Davies-Bouldin index is much more successful when using Euclidean distances extent. In the case of a Ward's method, successful is higher by 15, 63 %.

Key words: clustering, evaluating of clustering, methods, number of clusters

JEL Code: C 38, C 40

Introduction

The main aim of cluster analysis is the classification of objects, see (Gan et al 2007). There are various methods and and various distance measures to do that. These methods can be categorized according to various criteria see e.g. (Gan, 2007; Rezankova et al., 2009, Löster, 2014c).

Traditional methods are well developed and they are applied in many software products. Very important are the measures of similarities, resp. the distance measures. There are a lot of distance measures and in the practice they are combined with various clustering methods, see e.g. (Gan, 2007; Rezankova et al. 2009; Löster, 2014a; Löster, 2014b; Löster, 2015). Very frequently used is the Euclidean distance measure, see (Löster, 2014a; Löster, 2015). In the context of this article we will consider Euclidean and Mahalanobis distance measures. We will examine which results are achieved when we determine the number of clusters together with the various methods of clustering using different coefficients. For example RMSSTD, CHF, Davies-Bouldin, PTS and Dunn´s coefficient. Cluster analysis is very often used statistical method, see e.g. (Halkidi et al., 2001; Meloun, 2005; Löster 2012, 2014a, 2014b; Löster et al., 2015; Rezankova et al., 2013, Stankovičová at al, 2007). 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; Želinský et al, 2012). Other demographic variables, which are very often used in cluster analysis, are described in (Megyesiova, et al. 2011, 2012, Šimpach, 2012).

1 Clustering methods

Among the best known clustering methods can be included the nearest neighbour method, furthest neighbour method, centroid method, the average distance and also the Ward's method, see (Gan, et al., 2007). These methods are included for example in the SYSTAT software, which we used in this paper to evaluation of coefficients.

2 Distance measures

Euclidean distance is the most frequently used measure of distance, see (Gan, et al., 2007). It represents the length of the hypotenuse of a right-angled triangle. The calculation of the Euclidean distance measure of *i*th and *j*th object is based on the Pythagorean Theorem according to the formula

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

Mahalanobis distance, unlike the other distance measures such as Euclidean distance, Minkowski distance, etc, which are described for example in (Gan, 2007), eliminates the problem that arises when using non-standardized data, which may cause the differences between clusters, due to differences of measurement units. This distance measure is also applicable even in case that the individual variables are interdependent.

Mahalanobis distance between objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined by formula:

$$D_{Ma} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} S^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \quad (2)$$

where S is the covariance matrix.

3 Coefficients for determining the optimal number of clusters

Among the best known coefficients for determining the optimal number of clusters can be included CHF, PTS, RMSSTD, Davies-Bouldin (D-B) and Dunn´s coefficients, see (Gan, et al., 2007). These coefficients are included for example in the SYSTAT software and the researcher has also the option to apply these coefficients to determine the optimal number of clusters in connection with different methods and different distance measures.

4 Real data sets

To evaluate the coefficients for determining the optimal number of clusters in combination with different clustering methods and with different distances were used a total of 32 real data sets that originate from a known database *The UCI Machine Learning Repository,* see *https://archive.ics.uci.edu/ml/datasets.html*. This database includes various data files that have previously known number of clusters, and so the evaluation of the factors and is possible. These are the following data files: *Wine, Iris, Abalone, Cardiotocography, German Credit Data, Banknote Authentication, Blood Transfusion Service Center, Climate Model Simulation Crashes, Connectionist Bench (Sonar, Mines vs. Rocks), Ecoli, Echocardiogram, Energy Efficiency, Fertility, Haberman's Survival, Indian Liver Patient, Connectionist Bench (Vowel Recognition - Deterding Data), Ionosphere, Musk (Version 1), Parkinson Speach, Pima Indians Diabetes, QSAR Biodegradation, QSAR Biodegradation NV 1, QSAR Biodegradation NV 2, Seeds, Statlog (Vehicle Silhouettes) a+b, Statlog (Vehicle Silhouettes) a+g, Vertebral Column, Wall-Following Robot Navigation Data, Wholesale Customers, Susy NV 1, Susy NV 2 and Susy NV 3.*

5 Results

Based on a combination of different distance measures and different clustering methods were obtained different results of optimal number of clusters which provide individual coefficients. There are the number of cases (in %) in which the individual coefficients correctly determine the number of clusters using various clustering methods in combination with a Euclidean distance measure in table 1.

It shows for example, that the best results were obtained by using nearest neighbour method using Dunn´s coefficient. The success in determining the optimal number of clusters was 59, 38%. Coefficient RMSSTD can´t be used in combination with any method, because it´s succes did not exceed 20%.

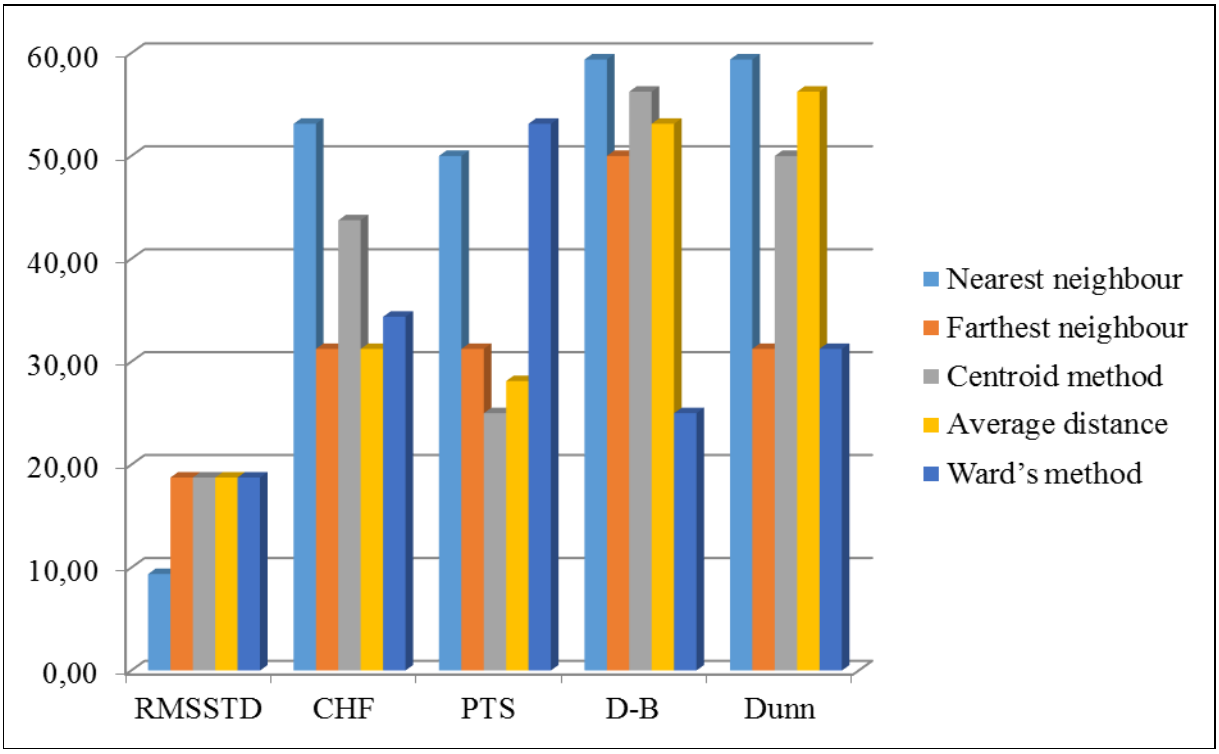
Tab. 1: Number of correctly set clusters (in %) – Euclidean distance measure

Methods/coefficients	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	9,38	53,13	50,00	59,38	59,38
Farthest neighbour	18,75	31,25	31,25	50,00	31,25
Centroid method	18,75	43,75	25,00	56,25	50,00
Average distance	18,75	31,25	28,13	53,13	56,25
Ward's method	18,75	34,38	53,13	25,00	31,25

Source: our calculation

Graphic representation of the success of the individual coefficients is shown in Figure 1. The most successful coefficients using Euclidean distance measure are Davies-Bouldin´s and Dunn´s index.

Fig. 1: The success of coefficients – Euclidean distance measure



Source: our calculation

There are the number of cases in which the individual coefficients correctly determine the number of clusters using various clustering methods in combination with a Mahalanobis distance measure in table 2.

It shows for example, that the best results were achieved using the method Centroid method using Davies-Bouldin coefficient. The success in determining the optimal number of clusters was 68, 75 %. Coefficient RMSSTD can´t be used in combination with any method again.

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

Methods/coefficients	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	6,25	50,00	46,88	56,25	46,88
Farthest neighbour	21,88	37,50	40,63	37,50	37,50
Centroid method	9,38	59,38	50,00	68,75	37,50
Average distance	9,38	53,13	46,88	65,63	53,13
Ward's method	28,13	50,00	37,50	9,38	59,38

Source: our calculation

Graphic representation of the success of the individual coefficients is shown in Figure 2. The most successful coefficient in using Mahalanobis can be considered Davies-Bouldin´s index.

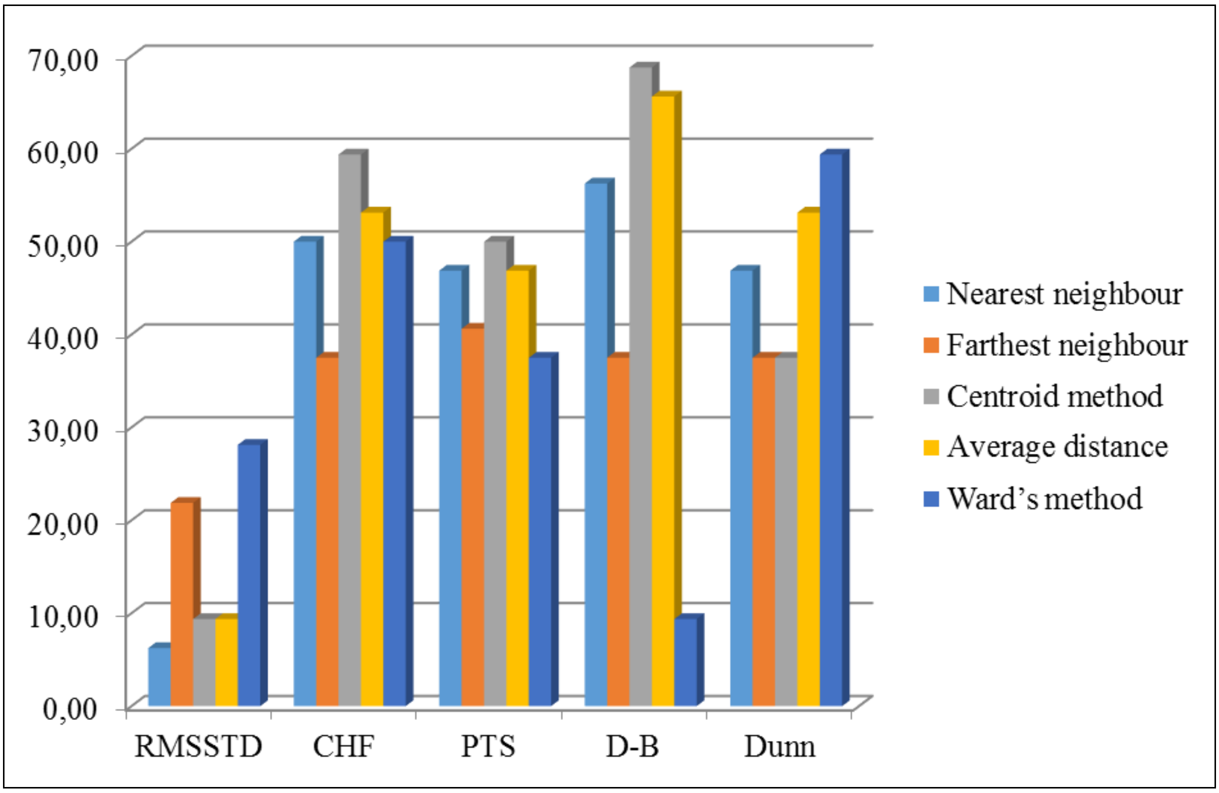
Tab. 3: Differences in the success rate (in %)

Methods/coefficients	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	**3,13**	**3,13**	**3,13**	**3,13**	**12,50**
Farthest neighbour	-3,13	-6,25	-9,38	**12,50**	-6,25
Centroid method	**9,38**	-15,63	-25,00	-12,50	12,50
Average distance	**9,38**	-21,88	-18,75	-12,50	**3,13**
Ward's method	-9,38	-15,63	**15,63**	**15,63**	-28,13

Source: our calculation

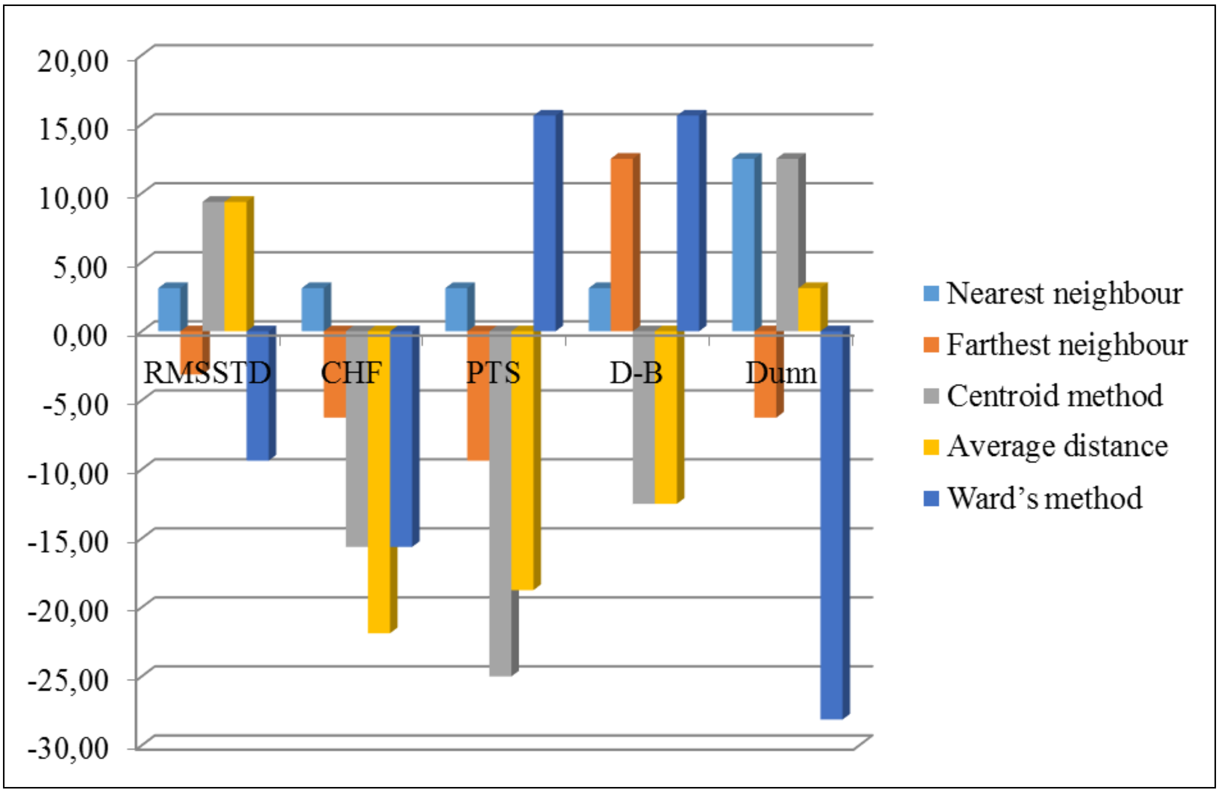
In Table 3, there are differences in the success rate of the coefficients using the Euclidean and Mahalanobis distance measure. It shows for example, that using of the Davies-Bouldin´s coefficient is better by using a Euclidean distance measure, while using CHF, PTS coefficient and coefficient is better by using Mahalanobis distance measure.

Fig. 2: The success of coefficients – Mahalanobis distance measure



Source: our calculation

Fig. 3: Difference in the success of coefficients (in %)



Source: our calculation

Graphical representation of the differences in the success of individual coefficients using Euclidean distance measure and Mahalanobis distance measure can be seen from Figure 3. The greatest difference is achieved by using of Dunn´s coefficient and its success, using Mahalanobis distance measure is almost by 30% higher.

Conclusion

Cluster analysis is a multivariate statistical method that is used to classify objects into clusters. There are many methods of clustering and there are many distance measures between objects in current literature. The combinations of different methods and different distance measures give different results. The current literature does not address the different combinations and there is nowhere stated that the combination is successful.

Part of the cluster analysis is usually also determining the optimal number of clusters in which individual objects are classified. Even in this case, there are many coefficients that can be used for this task. The main aim of this paper is on the 32 real datasets find suitable combinations that deliver the best results. We compared five methods of clustering coefficients and five coefficients to determine the optimal number of clusters. On the basis of different combinations we compared success of these clustering methods clustering in connection with Euclidean and Mahalanobis distance measures. These two measures were chosen because the first of them is very often used and the second of them eliminates a potential problem with correlations of variables that characterize the individual objects.

Based on these results, it was found that it is not possible to say which of distance measures is clearly more successful. It is always necessary to evaluate the combination of clustering methods, distance measures and the coefficient. Generally, when we compare the these two distance measures, we can say, that better results were obtained in more cases by using Mahalanobis distance measure. When we used Mahalanobis distance measure, we obtained the best results in determining the optimal number of clusters by using the Davies-Bouldin´s index, whose percentage success rate was 68, 75 % in connection with Average distance method and 65, 63 % in connection with Centroid Method. Conversely, when we used the Euclidean distance measure, the best results that we obtained were by using nearest neighbor method in connection with the Davies-Bouldin´s and Dunn´s index. The succes rate was in both case 59, 38 %.

Acknowledgment

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

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.

Löster, T. (2012). *Nerovnosti mezi regiony České republiky u podnikatelské sféry z hlediska trhu práce*. In: Nerovnosť a chudoba v Európskej únii a na Slovensku. [online] Herlany, 26.09.2012. Košice: Ekonomická fakulta TU, 2012, s. 123–130. ISBN 978-80-553-1225-5

Löster, T. (2014a). *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.

Löster, T. (2014b). *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.

Löster, T. (2014c). *Metody shlukové analýzy a jejich hodnocení.* 1. vyd. Slaný: Melandrium, 2014. 132 s. ISBN 978-80-86175-88-1.

Löster, T., Cséfalvaiová, K. (2015). Determining the number of clusters in the analysis of the consumer rationality. *Management and Engineering*, 2015, roč. XXIII, č. 1, s. 285–291. ISSN 1310-3946.

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*, 1. vydání, Academia, Praha, 2005.

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

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

Šimpach, 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).

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

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).

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