Information Technology and Management Science
doi: 10.2478/itms-2013-0013
2013 / 16

Clustering Algorithm for Travel Distance Analysis

Nadezda Zenina[1], Arkady Borisov[2], [1-2] *Riga Technical University*

***Abstract* – An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. The aim of the paper is to determine a number of clusters with a distinctive breaking point (elbow), calculating variance ratio criterion (VRC) by Calinski and Harabasz and J-index in order to check robustness of cluster solutions. Agglomerative hierarchical clustering was used to group a data set that is characterized by a complex structure, which makes it difficult to identify a structure of homogeneous groups. Stability of cluster solutions was performed by using different similarity measures and reordering cases in the dataset.**

***Keywords* – Agglomerative hierarchical clustering, distinctive breaking point (elbow), J index, variance ratio criterion**

I. Introduction

Cluster analysis identifies homogenous groups of clusters of cases without any prior information about the real classification. Partitioning, density-based, grid-based, and hierarchical algorithms are main groups of cluster analysis. The complexity of cluster analysis application is that how to verify stability of cluster solutions. A variety of methods are used to estimate and evaluate the number of clusters: cross-validation, penalized likelihood estimation, bootstrap based on Anova model [15], and finding the knee of an error curve [4].

Extensive comparative study was carried out in [13] with an aim to find the optimal number of clusters, comparing 30 methods for hierarchical clustering algorithms on well-separated data. According to their work, Calinski and Harabasz index and J-index are the most effective methods to determine the stability of cluster solutions.

The main goal of this paper is to verify stability and validity of cluster solution determining optimal number of clusters based on distinctive breaking point (elbow), Calinski and Harabasz index, J-index. Different similarity measures – Euclidean distance and Manhattan distance – were used for this purpose.

II. Hierarchical Clustering

Hierarchical clustering algorithms consist of the following steps:

1. choosing a hierarchical clustering technique;
2. selecting a measure of similarity;
3. selecting a Linkage Method;
4. data normalization;
5. representation of cluster results.

Further each step of clustering algorithm is described in detail.

A. Choosing a Hierarchical Clustering Technique

Hierarchical clustering algorithms are divided into two groups: agglomerative and divisive ones. In agglomerative clustering, each case starts in its own cluster and in the next step the two most closely located cases are merged till all cases are joined into a single cluster. In divisive clustering, all cases are located in one cluster and further are subdivided into clusters until all cases are located in their own clusters. In this study, agglomerative clustering was chosen for the analysis based on literature review [7], [8] and to determine a number of clusters for a data set, the following steps were performed:

1. Start with one *n* cluster.
2. Find the most similar clusters with cases located closely to each other and merge them into one cluster.
3. Repeat Step 2 until the number of clusters becomes one [6].

B. Selecting a Measure of Similarity

Two distance measures were considered. Euclidean distance (1) was chosen as a measure to express similarity between pairs as the shortest path between two samples (Fig. 1).

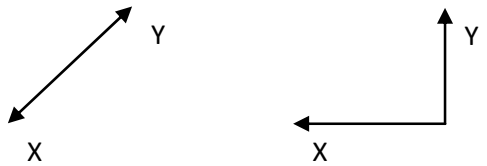
$$d_{ij} = \sqrt{\sum_{n=1}^{k} (x_{in} - x_{jn})^2} . \quad (1)$$

where xi, yj – points in Euclidean space.

Manhattan distance (2) was chosen because the analysed data set contains discrete data [16] and calculates distances along each dimension (i.e., “walking round the block”).

$$d_{ij} = \sum_{n=1}^{k} \left| x_{in} - x_{jn} \right| . \quad (2)$$

where xj, yj – points in n-space.



a) b)

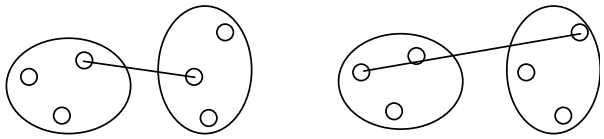
Fig. 1. Distance. a) Euclidean distance; b) Manhattan distance

C. Selecting a Linkage Method

Linkage methods calculate the distance from a cluster centre to a certain case. The most popular agglomerative clustering procedures include the followings linkage methods [11]:

1. Single linkage. The distance between two clusters is calculated as the shortest distance between any two cases in the two clusters.

2. Complete linkage. The distance between two clusters is calculated as the longest distance between any two cases in the two clusters.
3. Average linkage. The distance between two clusters is calculated as the average distance between all pairs of the two clusters.



a) b)

Fig. 2. Linkage methods. a) Single linkage; b) Complete linkage

Each linkage method could lead to different results for the same initial data. Single linkage method is based on minimum distances and tends to organize one large cluster with the other clusters containing only one or a few objects.

D. Data Normalization

To reduce the influence of variables on the clustering solution, z-score was used for data normalization (each variable should have a mean of 0 and a standard deviation of 1).

E. Representation of Cluster Results

Result of hierarchical clustering is a dendrogram (Fig. 1) that represents each merge at the similarity between the two merged groups [15].

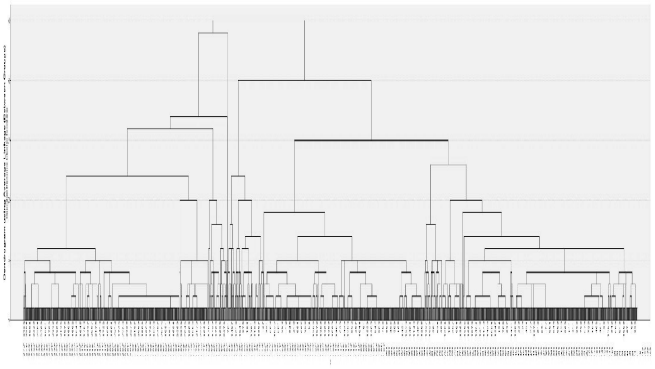


Fig. 3. Representation of cluster results – a dendrogram

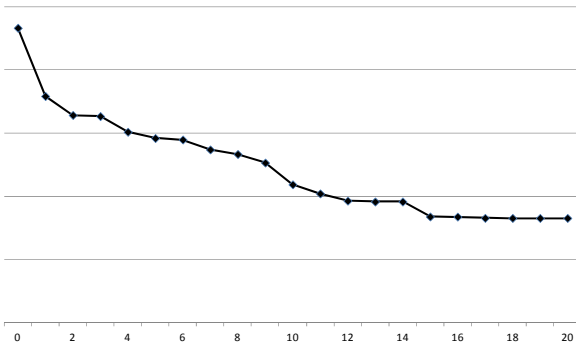
III. Stability and Validity of Cluster Results

Robustness of cluster is an important task in the clustering analysis to choose representative cluster solutions. To check cluster solutions for stability, the order of cases was changed. Cases were sorted randomly and for different order of cases, different cluster solutions were received. As a result, the solution with the highest goodness of fit was selected for the analysis. Multiple runs with different clustering procedures, algorithms or distance measures were performed. Euclidean distance and Manhattan distance similarity measures with single linkage were considered for this purpose.

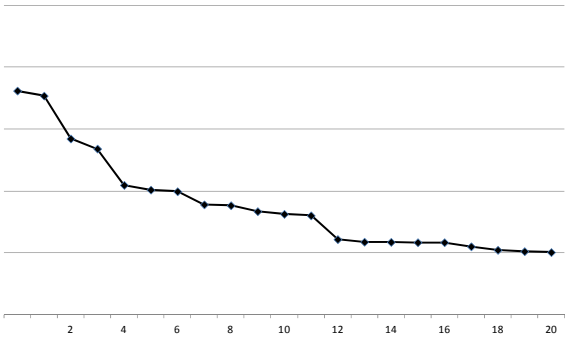
To check cluster validity, the appropriate number of clusters was determined by a distinctive breaking point, Calinski and Harabasz index and J-index.

A. Distinctive Breaking Point (Elbow)

Hierarchical procedures provide information that allows identifying the gaps that define logical clusters based on the output (Fig. 4). Sometimes it is difficult to identify where the break actually occurs [11].



a) Euclidean distance



b) Manhattan distance

Fig. 4. Determination of the optimal number of clusters

According to the study presented in [2], Calinski and Harabasz index is the most effective method to determine the optimal number of clusters, followed by Duda and Hart method (J index). Variance ratio criterion was used to calculate a number of clusters by Calinski and Harabasz index.

B. Variance Ratio Criterion (VRC) by Calinski and Harabasz

Variance ratio criterion introduced by Calinski and Harabasz [10] is a widely used criterion that computes ratio of between and within-cluster sums of squares for k clusters. The optimal solution of this criterion is the number of clusters that maximises the value of the variance criterion (3) and minimises ω_k value (4).

$$VRC_k = \frac{B_k}{k-1} \cdot \frac{n-k}{W_k} \tag{3}$$

where VRC_k – the variance ratio criterion, k – the number of clusters, B_k – the overall between-cluster variation, W_k – the overall within-cluster variation with respect to all clustering variables, n – data objects.

Value ω_k should be computed for each cluster solution to determine the optimal or suitable number of clusters.

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}). \quad (4)$$

The main limitation factor of Calinski and Harabasz index is that the number of clusters cannot be less than three, because the number of cluster is calculated based on the previous cluster information (VRC_{k-1}).

C. The J-index

The J-index proposed by Duda and Hart [11] compares the within-cluster sum of squared distance with the sum of within-cluster sum of squared distances and decides whether cluster should partitioned into two clusters. The hypothesis that cluster could be subdivided is rejected if DH value (5) is more than a standard normal quantile. In this study, z_{1-a} value is equal to 3.2 by [18].

$$DH = \left(-\frac{W_2}{W_1} + 1 - \frac{2}{\pi p}\right)\left(\frac{2(1-8/\pi^2 p)}{np}\right)^{-1/2} > z_{1-\alpha} \quad (5)$$

where p – the number of variables, n – the number of objects in the studied cluster and z_{1-a} – the standard normal quantile.

IV. Experiments

Two types of similarity measures (Euclidean distance and Manhattan distance) of hierarchical cluster algorithms, variance ratio criterion by Calinski and Harabasz and J-index by Duda and Hart were used to calculate an optimal number of clusters and check validity of cluster solutions.

Initial data for experiments were preprocessed with a purpose to clean the noisy data and to convert the data into a proper format. The analysed dataset after preprocessing includes ten variables and 2000 cases.

Variance ratio criterion and J-index were calculated with SPSS 16.0 and SAS statistical packages.

A. Distinctive Breaking Point (Elbow)

The elbow points shown in Fig. 4 suggest that it is not a clear elbow with rapid growth of distance indicating an appropriate number of clusters. A nine-cluster solution is the optimal choice for Euclidean distance and eleven clusters –for Manhattan distance based on numerical results.

B. Variance Ratio Criterion and J-index

Variance ratio criterion by Calinski and Harabasz was calculated for a number of clusters in the range from two to fifteen for Euclidean and Manhattan distance similarity measures.

The optimal number of cluster was a solution with the smallest ωk value. For Euclidean distance the suitable number of clusters was ten and for Manhattan distance it was twelve.

TABLE I

VARIANCE RATIO CRITERION FOR EUCLIDEAN AN MANHATTAN DISTANCES

Number of clusters	Similarity measure: Euclidean distance		Similarity measure: Manhattan distance	
	Variance ratio criterion	ω_k	Variance ratio criterion	ω_k
2	204.4		127.9	
3	104.97	71	68.07	68.3
4	76.54	39.69	76.54	-6.36
5	87.8	-28.41	78.65	-3.75
6	70.65	20.12	77.01	-8.2
7	73.62	-2.51	67.17	5.85
8	74.08	-9.25	63.18	7.31
9	65.29	925.68	66.5	-2.04
10	982.18	**-1012.01**	67.78	-7.68
11	887.06	20.24	61.38	757.2
12	812.18	17.46	812.18	**-806.22**
13	754.76	15.94	756.76	-1.03
14	713.28	852.44	700.31	21.99
15	1524.24	494.57	665.85	-7.78

J-index was calculated for the number of clusters in the range from two to fifteen for Euclidean distance. Eleven clusters represented the optimal solution with values 0.50/83.88, where the first number – the sum of squared error within the group and the second one – the sum of squared error in the two subgroups.

Results of the determination of number of clusters for hierarchical clustering algorithm solutions showed that the optimal number of clusters was eleven because two types of indices showed the same results.

TABLE II

VALUES OF THE THREE INDICES FOR THE DETERMINATION OF NUMBER OF CLUSTERS

Index		Number of clusters			
		9	10	11	12
"Elbow"	Euclidean distance	**0.096**	0.102	0.109	0.127
	Manhattan distance	0.083	0.081	**0.80**	0.61
Calinski and Harabasz	Euclidean distance	925.68	**-1012.01**	20.24	17.46
	Manhattan distance	-2.04	-7.68	757.2	**-806.22**
Duda and Hart	Euclidean distance	0.81/ 20.99	1.00/ 0.07	**0.50/ 83.88**	0.99/ 0.03

V. Conclusion

In this study, the application of agglomerative hierarchical clustering algorithm was presented. Distinctive breaking point (elbow), variance ratio criterion and J-index were calculated to determine the optimal number of clusters and to check the hierarchical clustering solutions for validity and stability. Before clustering all analysed data were pre-processed to clean noise.

Results of elbow point did not reach clear elbow in the plot for the considered distances, the number of suitable clusters was determined from numerical calculations.

The calculation of number of clusters by variance ratio criterion and J-index showed that the optimal number of clusters was in the range from nine to twelve. Application of distinctive breaking point with Manhattan distance and J-index

with Euclidean distance showed equal results, the number of cluster was equal to eleven.

REFERENCES

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, 1973

[2] G. W. Milligan and M. C. Cooper. *An examination of procedures for determining the number of clusters in a data set*. Psychometrica: 50, 1985, pp. 159-179.

[3] A.Ben-Hur and I.Guyon. *Detecting stable clusters using principal component analysis. In Functional Genomics:Methods and Protocols.* M.J.Brownstein and A.Kohodursky (eds.) Humana press, 2003, pp. 159-182.

[4] S. Salvador and P. Chan. *Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. 16th IEEE International Conference on Tools with Artificial Intelligence*, 2004, pp. 576 – 584.

[5] R.Tibshirani and G. Walther and D.Botstein and P.Brown. Cluster validation by prediction strength. Stanford Technical Report, Department of Statistics, Stanford University, USA, 2001.

[6] N.Rajalingam and K.Ranjini. *Hierarchical Clustering Algorithm - A Comparative Study. International Journal of Computer Applications*. Volume 19, No.3, April 2011.

[7] J. Anable. Complacent Car Addicts' or 'Aspiring Environmentalists'? Identifying travel behavior segments using attitude theory. *Transport Policy* 12, 2005, pp. 65–78.

[8] P. Hagel and R. Shaw. The Influence of Delivery Mode on Consumer Choice of University. European Advances in Consumer Research, Volume 8, 2008.

[9] S. Limbourg and B. Jourquin. Rail-Road terminal locations: aggregation errors and best potential locations on large networks. *EJTIR*, Vol 7, no. 4, 2007, pp. 317-334.

[10] R. B. Calinski and J. Harabasz, A dendrite method for cluster analysis, Comm. in Statistics, Vol 3, 1974, pp. 1–27.

[11] E. Mooi and M. Sarstedt. A Concise Guide to Market Research. The Process, Data, and Methods Using IBM SPSS Statistics, 2011.

[12] I. M. G. Dresen and T. Boe and J. Huesing and M. Neuhaeuser and K.-H. Joeckel1. New resampling method for evaluating stability of clusters. *BMC Bioinformatics*, 2008.

[13] G. W. Milligan and M. C. Cooper. A study of variable standardization. *Journal of Classification*, pp. 181–204, 1988.

[14] J. Smith and M. Saito, "Creating Land-Use Scenarios by Cluster Analysis for Regional Land-Use and Transportation Sketch Planning", *Journal of Transportation and Statistics*, vol. 04, no. 01, paper 03. [Online]. Available: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/journal_of_transportation_and_statistics/volume_04_number_01/paper_03/index.html . [Accessed June 15, 2013].

[15] Y. Mingjin, "Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion", Dr. thesis. [Online]. Available: http://scholar.lib.vt.edu/theses/available/etd-12062005-153906/unrestricted/Proposal-Face.pdf. [Accessed July 8, 2013].

[16] ArcGIS home page. [Online]. Available: http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000005000000. [Accessed May 29, 2013].

Nadezda Zenina is a postgraduate student at the Faculty of Computer Science, Riga Technical University (Latvia). She received her MSc. degree from Riga Technical University, the Department of Modelling and Simulation of in 2006.

Since 2007 she has been working at Solvers Ltd (Latvia) as a Transportation and Modelling Engineer. Her skills cover the fields of transportation engineering, transportation planning and transportation modelling. Research areas include artificial neural systems, data mining methods – learning trees, multinomial logit and discriminant analysis, cluster analysis, classification, traffic modelling, transportation sustainability.

E-mail: nadezdat@gmail.com

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and Dr.habil.sc.comp. degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

He is a Professor of Computer Science at the Faculty of Computer Science, Riga Technical University (Latvia). The research areas include artificial intelligence, decision support systems, fuzzy set theory and its applications and artificial neural systems. He has 235 publications in the field.

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).

E-mail: arkadijs.borisovs@cs.rtu.lv.

Nadežda Zeņina, Arkādijs Borisovs. Klasterizācijas algoritmi ceļošanas distances analīzei

Klasteru analīzi pielieto, lai identificētu homogēnas novērojumu grupas, nezinot informāciju par „īsto" datu sadalījumu. Klasteru analīzes pielietošanas sarežģītība ir saistīta ar to, kā novērtēt klasteru risinājumu stabilitāti un noteikt nepieciešamu klasteru skaitu datu sadalīšanai/sagrupēšanai. Darba mērķis - noteikt klasteru skaitu pēc „elkoņa" metodes (tiek aprēķināta starpība starp apvienošanas līmeņiem, klasteru skaits tiek noteikts pēc dendrogrammas), Calinski un Harabasz kritēriju un J-indeksa pamata, lai novērtētu klastera risinājuma stabilitāti. Aglomeratīvā hierarhiskā klasterizācija tika izmantota, lai sagrupētu datus ar sarežģītu struktūru un identificētu homogēnas grupas. Klasteru risinājumu stabilitāte pārbaudīta, pielietojot dažādus attāluma mērus un mainot objektu kārtību datu izlasē. Klasterizācijas rezultāti, pielietojot „elkoņa" metodi kopā ar Eiklīda un Manhhatan distanci, neuzrādīja spilgti izteiktu izlieci dendrogrammā, un klasteru skaits tika noteikts, pamatojoties uz skaitliskiem aprēķiniem. Calinski un Harabasz dispersiju samēra kritērijs tiek rēķināts kā starpklasteru distances matricas attiecība pret iekšklasteru distances matricu. Kritērijs tika aprēķināts diviem līdzības mēriem – Eiklīda distancei un Manhhatan distancei, un klasteru skaitam no diviem līdz piecpadsmit. Analīzes rezultāti uzrādīja, ka, pielietojot Eiklīda līdzības mēru, optimālais klasteru skaits ir vienāds ar desmit, Manhattan distances gadījumā – divpadsmit klasteru. J-indekss salīdzina iekšklasteru attālumu kvadrātu summu, lai noteiktu, vai ir iespējams klasteru sadalīt divās daļās. Hipotēze par klasteru sadalīšanu tiek noraidīta, ja DH vērtība ir lielāka par standarta normālo sadalījuma kvantīli. Optimālais klasteru skaits pēc J-indeksa sastādīja vienpadsmit klasteru. Klasteru skaita noteikšana, pamatojoties uz dendrogrammu, pielietojot Manhattan distanci, kā arī J-indekss uzrādīja līdzīgus rezultātus, optimālais klasteru skaits bija vienpadsmit.

Надежда Зенина, Аркадий Борисов. Алгоритмы кластеризации в анализе расстояния путешествия

Кластерный анализ применяется для идентификации однородных групп наблюдений без исходной информации о «настоящем» разделении данных. Сложность применения кластерного анализа заключается в том, как оценить стабильность кластерных решений и определить, на какое количество кластеров необходимо разбить выборку. Цель работы - определить количество кластеров на основе разницы между уровнями объединения (определение количества кластеров на основе дендрограммы), максимального значения показателя псевдо-F-статистики Calinski и Harabasz, и J-индекса для оценки работоспособности кластерного решения. Аггломеративная иерархическая кластеризация была применена для группировки данных, характеризующихся сложной структурой, для идентификации однородных групп. Стабильность кластерных решений проверена, применяя различные меры схожести объектов и меняя порядок наблюдений в базе данных. Результаты на основе дендрограммы не показали ярко выраженного изгиба на дендрограмме, применяя Эвклидово и Манхеттенское расстояния, и количество кластеров было определено на основе численных расчетов. Критерий соотношения дисперсий (Calinski and Harabasz) рассчитывается как соотношение матрицы межкластерных расстояний к матрице внутрикластерных расстояний. Критерий был рассчитан на основе двух мер схожести между объектами - Эвклидово расстояние и Манхеттенское расстояние, для количества кластеров от двух до пятнадцати. Результаты анализа показали, что применяя Эвклидово расстояние оптимальное количество кластеров равно десяти, в случае Манхеттенского расстояния – двенадцать кластеров. J-индекс сравнивает сумму квадратов внутрикластерных расстояний с целью определить, можно ли кластер разбить на два. Гипотеза о том, что кластер может быть разделен, отклоняется, если значение DH больше, чем квантиль стандартного нормального распределения. Оптимальное количество кластеров по J-индексу составило одиннадцать кластеров. Определение количества кластеров на основе дендрограммы с использованием Манхеттенского расстояния и J-индекс показали схожие результаты, оптимальное количество кластеров составило одиннадцать.