SURVEY OF THE CLUSTERING METHODS

Gergely Endre Tóth – Gábor Fodor – Raymund Pardede –
Loránd Lehel Tóth – György András Jeney – Gábor Hosszú

1. SUMMARY

This article presents the fundamental parts and classes of the cluster analysis, and shows their advantages and disadvantages along with the applied methods. The article also elaborates each cluster analysis method and explains how to choose between the available methods. Moreover, the technique to analyse the result is also presented. The last part of this article covers, how the cluster analysis methods can be applied for data mining and knowledge discovery.

2. OVERVIEW

In this decade, information becomes the primary needs of the human being, and that information is obtained by processing the available data. Most of the time, the available data is very large and not easy to evaluate because of its complexity. The stock exchange results may be a good example of the complex data, which therefore nowadays there are many tools and computer applications are developed to process the data to produce the result to be used as an information for the users.

The needs for processing data apply not only to certain group of people (e.g. the stock exchange traders), but also to anyone who needs any kind of information. As an example, in the daily life when someone read, listen, or watch the news, the human brain processes the input data to distinguish between the necessary and the unnecessary information to digest. The process to discover this information is called *building* process, which includes the handling, analysing, and grouping the large amount of data [1] [2] [3]. In general, the overall method is called *data mining*. The main goal of the data mining method is to process and analyse data from different perspective and summarizing it into useful information.

The data mining method uses mathematical statistics approach to process all the input data. It is also important to have the data pre-processed before being used as an input for that method. The obtained result later can be used by the user to select what is necessary. As a practical example, when someone who is interested in the Hungarian water polo is looking for a match result on a news portal, the portal should automatically present the relevant information. When the data is large, it would take and effort to process it with the traditional method. Therefore, the data mining method should be used since it works more efficiently in those cases.

The data mining consists of several methods, and one of them is called *cluster analysis*. Cluster analysis classifies data objects based only on information found in the data that describes inherent structure of the data objects [1]. The purpose of the cluster analysis is that the object within a cluster be similar to another ad different from the objects in other clusters. The cluster analysis techniques are concerned with exploring data sets to assess whether or not they can be summarized meaningfully in terms of a relatively small number of groups or clusters of objects or individuals which resemble each other and which are different in some respects from individuals in other clusters [2]. The cluster analysis also contains various kind of algorithms. In most cases the cluster analysis algorithms need to be adjusted for solving the given problem. The reason is because each problem has its unique data and present different kind of real life problems. The used technique must be fitted to the given structure (or reverse) [4].

In the following chapters some fundamental techniques of the cluster analysis will be presented. The principles to select the suitable algorithm for resolving a unique data processing problem are also presented.

3. SIMILARITY AMONG DATA

In cluster analysis similarity and dissimilarity are important, because it makes groups by right of these. The attributes of the observation have to be quantified. Using these similarity (or dissimilarity) can be measured between the attributes of the observation. There are four types of attributes [1]: *Categorical* (with other term Qualitative) and *Numeric* (with other term Quantitative) type. The Categorical type refers to data if the values belonging to it can be sorted accordingly to category. It is noteworthy, that each value would be chosen from a set of non-overlapping categories. The Numeric data values are counts or numerical measurements. A Numeric data can be either continuous or discrete. The Categorical type includes the *Nominal* and the *Ordinal* data. A set of data is called Nominal, if the values are just different names, we can distinguish one object from another. Oppositely, the data is called Ordinal if the values belonging to it can be ranked. Note that counting or ordering the Ordinal data is possible; however, this kind of data cannot be measured. The Numeric type can be divided into the *Interval* and the *Ratio* data. In case of the Interval data, the addition and subtraction between objects can be defined. However, in case of Ratio data, multiplication and division between objects are also meaningful.

There are systematic methods, which are usable to quantify the data [1], however, the human interaction is also necessary. Quantification defines what methods can be used. For example, if we use nominal types of data, we can measure similarity and dissimilarity between the attributes. If we use numeric types we can measure similarity and dissimilarity with mathematical distance functions (for example Euclidean-, Manhattan distance etc.). In most cases the Euclidean distance can be applied, but in some cases other distance metrics are necessary. For example in heart sound analysis City Block distance (sum of absolute differences between coordinates of objects) has been used, because of the different units [5]. In some special cases, normalization has to be used to scale the different values into the same level. It is especially important when the magnitude of attributes is different.

Using quantified attributes the objects can be divided into groups by clustering. The goal is to capture the natural structure of the data. In several cases, cluster analysis is only a starting point for detecting unknown connections or structures in the given dataset [1]. With the created groups the large amounts of data can be processed by human.

The collection of clusters is generally referred to as a clustering, which have different types: *hierarchical* and *partitional* (with other term partitioning-based) clustering. In case of hierarchical clustering the clusters are permitted to have subclusters. In this case the clusters are organized as a tree. Each non-leaf node in the tree is the union of its subclusters. Moreover, the root of the hierarchical tree is a cluster, which contains all others. Oppositely, the partitional clustering is a division of the set of data objects into non-overlapping clusters such that each data object is in accurately one cluster. The hierarchical clustering can be viewed as a sequence of partitional clusterings [1].

4. PARTITIONAL CLUSTERING METHODS

The great advantage of partitional methods is their speed and easy implementation [2], however, the result is irrelevant in some cases. Moreover, the number of clusters must be determined before performing the algorithm [6]. Like other iteration methods the solution my highly depend on the initial starting conditions [1]. Thus the algorithm reaches local minima, instead of finding the global solution. The output can be validated using quality measures (e.g. sum of squared errors) [1] [2]. Performing several replicates beginning from different randomly selected initial conditions can decrease the chance to find a local optimum [6]. The one with the best quality measure can be accepted over all replicates. If we have existing knowledge about the initial dataset it can be used to verify the solution.

The K-means algorithm minimizes the sum of distances [6]. This algorithm takes $O(i \cdot K \cdot m \cdot n)$ time and $O((m+K) \cdot n)$ space, where m is the number of objects, K is the number of clusters, n is the dimension and i is the number of iterations. Several statistical analysis programs use built-in functions to implement the algorithm e.g. the Mathworks Matlab Statistical Toolbox [8], the SPSS [9], and the R software [10]. The K-means method has several improved versions [6], which optimize the determination of the initial conditions and the recalculation of the cluster centroids.

Another partitional method is the fuzzy based algorithm called FCM, which is very similar to K-means [1]. Unlike hard clustering, where each object belongs to exactly one cluster FCM calculates a set of coefficients to each point giving the degree of being in a given cluster [4] [10].

5. HIERARCHICAL CLUSTERING METHODS

The hierarchical methods are more resource-dependent in comparison with iterative methods. Their time and space complexity can be calculated as $O(m^2)$ [6]. The advantage of the hierarchical methods is that it is not necessary to know the exact number of clusters. The output of the method is a cluster tree or so-called dendrogram. It allows graphically to decide the level or scale of clustering. It represents a multilevel hierarchy as the function of the given distance (similarity) measure [6] [11]. If there is not any information about the number of clusters in a dataset, the hierarchical methods can be used. However, it cannot be applied to large datasets.

There are several varieties of hierarchical algorithms, each with its own data structure on which works efficiently [6]. Different chain methods can be used to find similarity between clusters and arbitrary distance measuring functions can be applied [1] [6]. Because of the lot of free parameters the analysis must be performed with different initial conditions. Several quality measures and existing knowledge about the structure of the dataset can be used to verify the results [11].

6. PREPROCESSING, DETERMINING THE NUMBER OF CLUSTERS

The clustering algorithms can be more efficient if we have existing knowledge about the structure of the dataset which helps to determine the initial parameters [1]. Moreover it can be used to verify the results. If the number of clusters are not known either hierarchical analysis methods or quality measures can be used to evaluate the created clusters.

The following example uses the average distance and a quality measure to determine the optimum number of clusters in a given dataset [12]. The quality measure ranges from [-1, +1]. Value 1 indicates points that are very far from neighboring clusters, 0 indicates points that are not distinctly in one cluster or another and -1 indicates points that are probably assigned to the wrong cluster, respectively. It measures how close each point in one cluster is to points in the neighboring clusters, and is defined as:

$$qm(i)=\frac{\min\{dist_b(i)\}-dist_w(i)}{\max\{dist_w(i),\min\{dist_b(i)\}\}}, \tag{1}$$

where $dist_w$ is the average distance from the i-th point to the other points in its own cluster, and $dist_b$ is the average distance from the i-th point to points in another cluster. Finally, the average value of qm is calculated (qm_{ave}).

The values of these measures can be seen in Figure 1 corresponding to different k cluster numbers. The K-means clustering algorithm was used. Figure 1 (A) shows the average distance between cluster elements in the function of k cluster numbers. Figure 1 (B) shows the quality measure defined in equation (1). In Figure 1 (A) the optimum point is where the slope of the curve rapidly decreases. In Figure 1 (B) the local maxima can be accepted.

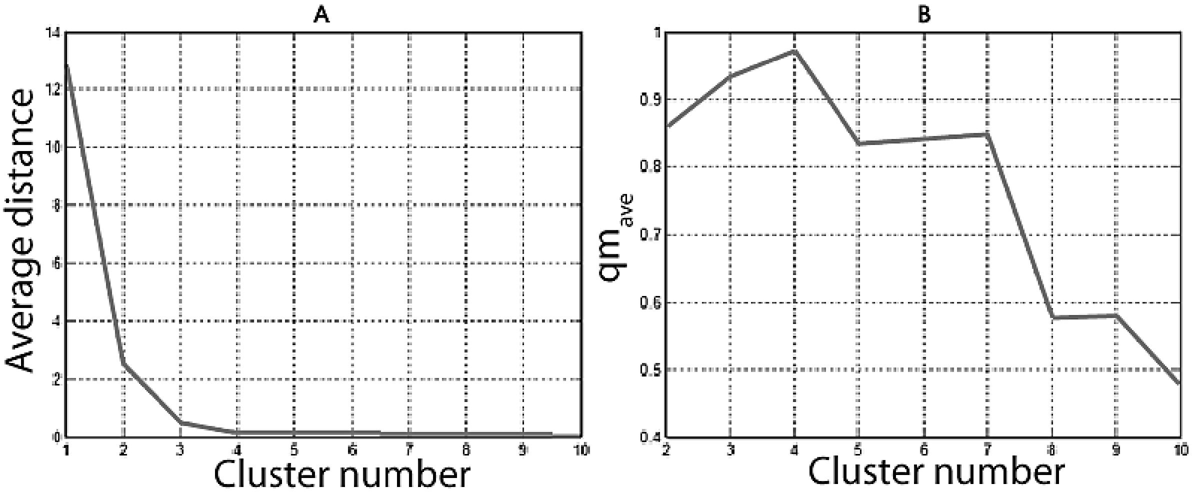

Figure 1.
(A) The average distance between cluster elements in the function of k cluster numbers.
(B) Shows the qm_{ave} quality measure value of each class.

7. SUMMARY

The article surveyed the cluster analysis and compared the advantages and disadvantages of the main classes of the cluster analysis methods. Some methods for determining the optimal number of the clusters were also presented. The useful representation tool, the dendrogram was introduced.

Moreover, the methods of interpretation and validation of clusters of data were also demonstrated. One of them is the so-called silhouette, which is a succinct graphical representation of how well each object lies within its cluster. The average silhouette width can be applied to select an optimal number of clusters.

8. ACKNOWLEDGEMENT

The work reported in the paper has been developed in the framework of the project “Talent care and cultivation in the scientific workshops of BME”. This project is supported by the grant TÁMOP - 4.2.2.B-10/1--2010-0009.

9. LIST OF REFERENCES

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar: *Introduction to Data Mining*, Addison Wesley, 1 edition (May 12, 2005), ISBN 978-0321321367

[2] Brian S. Everitt, Dr Sabine Landau, Dr Morven Leese, Dr Daniel Stahl: *Cluster Analysis*, Wiley; 5 edition (March 8, 2011), ISBN 978-0470749913

[3] Masoud Mohammadian: *Intelligent Agents for Data Mining and Information Retrieval*, IGI Global (February 2004), ISBN 1-59140-277-8

[4] Leonard Kaufman, Peter J. Rousseeuw: *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience; 1 edition (March 25, 2005), ISBN 0-47 1-73578-7

[5] G. Fodor, G. Hosszú, & F. Kovács: Model Improvement and Cluster Analysis of the Fetal First Heart Sounds, 2011. European IFMBE MBEC ‘11, 5th European Conference of the International Federation for Medical and Biological Engineering

[6] Guojun Gan, Chaoqun Ma, and Jianhong Wu: *Data Clustering Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, SIAM, Society for Industrial and Applied Mathematics (May 30, 2007), ISBN 978-0-898716-23-8

[7] János Abonyi, Balázs Feil: *Cluster Analysis for Data Mining and System Identification*, Birkhäuser Basel; 1 edition (August 17, 2007), ISBN 978-3-7643-7987-2

[8] Matlab Manual, http://www.mathworks.com/help/toolbox/stats/bq_679c.html, letöltve 2012.01.03-án

[9] Raynald Levesque and SPSS Inc.: *SPSS Programming and Data Management: A Guide for SPSS and SAS Users*, Fourth Edition (2007), SPSS Inc., ISBN 1568273908

[10]Francois Husson, Sébastien Lê, and Jérôme Pagès. *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC Computer Science & Data Analysis, 2010. ISBN 978-1-4398-3580-7

[11]H. Charles Romesburg: *Cluster analysis for researchers*, Lifetime Learning Pub (March 1984), ISBN 0-534-03248-6

[12]Rousseeuw, Peter J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, November 1987, pp. 53-65.