SOCIETY. INTEGRATION. EDUCATION
Proceedings of the International Scientific Conference. Volume V, May 24th -25th, 2019. 344-354

THE POSSIBILITIES OF CLUSTERING LEARNING METHODS IN STUDENT EDUCATION

Peter Grabusts
Rezekne Academy of Technologies, Latvia

Abstract. *Many educational courses operate with models that were previously available only in mathematics or other learning disciplines. As a possible solution, there could be the use of package IBM SPSS Statistics and Modeler in realization of different algorithms for IT studies. Series of research were carried out in order to demonstrate the suitability of the IBM SPSS for the purpose of visualization of various simulation models of some data mining disciplines – particularly cluster analysis. Students are very interested in modern data mining methods, such as artificial neural networks, fuzzy logic and clustering. Clustering methods are often undeservedly forgotten, although the implementation of their algorithms is relatively simple and can be implemented even for students. In the research part of the study the modelling capabilities in data mining studies, clustering algorithms and real examples are demonstrated.*
Keywords: *clustering, data analysis, modelling, simulation, SPSS, SPSS Modeler, learning.*

Introduction

Methods of data analysis and automatic processing are treated as knowledge discovery. That is why the notion of similarity is becoming more and more important in the context of intelligent data processing systems. It is frequently required to ascertain how the data are interrelated, how various data differ or agree with each other, and what the measure of their comparison is. Clustering methodology can be widely used in modeling, evaluation of different economic, financial and educational processes.

Nowadays there is a large amount of data in various fields of science, business, economics, etc. and there is a need to analyse them for better management of a particular industry. The goal of cluster analysis as one of the basic tasks of intellectual data analysis is to search for independent groups (clusters) and their characteristics in analytical data. Solving this problem allows for better understanding of data, since clustering can be practically used in any application area where data analysis is required.

The cluster analysis is based on the hypothesis of compactness. It is assumed that the elements of the training set in the feature room are compact. The main task is to formally describe these formations. All clustering algorithms have common parameters, the choice of which also characterizes clustering efficiency.

© Rēzeknes Tehnoloģiju akadēmija, 2019
http://dx.doi.org/10.17770/sie2019vol5.3723

The most important parameters characterizing clustering are: metrics (the distance of cluster elements to the cluster center), the number of clusters k.

The aim of the article is to show SPSS Modeler suitability for the purpose of visualizing simulation models of various data analysis disciplines. To reach the aim, the following research tasks have been set: identification of SPSS Modeler possibilities for clustering algorithms; demonstrate visualization models on the basis of examples; showing the possibilities of clustering algorithms operation for training purposes. Common research methods are used in this research: descriptive research method, statistical method and mathematical modelling.

Application of clustering methods in data analysis

The issue of "How to organize observable data in reviewed structures?" is a topical issue in various research areas. There is an opinion that unlike many other statistical procedures, in most cases, cluster analysis methods are used when there are no hypotheses regarding to classes, but data collection is still in progress. Cluster analysis methods allow to split exploratory objects into groups of "similar" objects called clusters (Kaufman & Rousseau, 2005; Aggarwal & Reddy, 2013; Wierzchon & Klopotek, 2018). The essence of clustering is depicted in Figure 1, where the two-dimensional space objects are conditionally divided into 5 clusters.

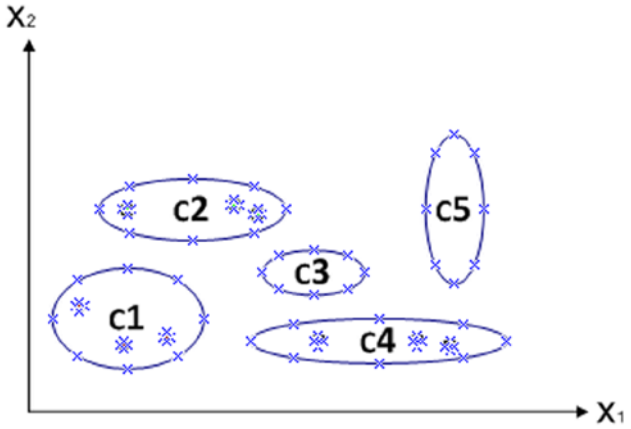


Figure 1 ***An example of a two-dimensional object space division into clusters***

Clustering differs from the classification by the fact that there is no need to separate a changeable group for analysis in the clustering process. From this point of view, clustering is treated as "non-teacher training" and is used in the initial phase of the research (Xu & Wunch, 2009).

The cluster analysis is characterized by two features that distinguish it from other methods:

1) the result depends on the nature of the objects or their attributes, i.e. they can be uniquely determined objects or objects with a fuzzy description;
2) the result depends on the possible relationship between the cluster and the objects in the clusters, i.e., the possibility of belonging the object to several clusters and the determination of the ownership of the object (strong or fuzzy belonging) must be taken into account.

Taking into account the important role of clustering in data analysis, the concept of object belonging was generalized to the function of classes that determines the class objects belonging to a particular class.

Two types of classes characterizing functions are distinguished:

1) discrete function that accepts one of the two possible values - belongs to / does not belong to the class (classical clusterization)
2) a function that accepts values from the interval [0,1]. The closer the values of the function to 1, the "more" the object belongs to a particular class (fuzzy clustering).

Clustering algorithms are mainly intended for the processing of multidimensional data samples, when the data is given in the form of the table "object-property". They allow you to group objects in defined groups, in which objects are related to each other according to a particular rule. It does not matter how the following groups are called - taxons, clusters, classes, the main thing that they accurately represent the properties of these objects. After clustering, other intelligent data analysis methods use data for further analysis in order to find out the nature of the acquired regularities and the possibilities for future use (Han et al., 2001).

Clustering is commonly used in the data processing as a first step of analysis. It identifies similar data groups that can later be used to explore the interrelationships of data (Gan et al., 2007; Han et al., 2001). The cluster analysis process formally consists of the following steps:

- collection of data necessary for analysis;
- determination of cluster characterizing sizes and boundaries;
- grouping data in clusters;
- class hierarchy determination and analysis of results.

The K-Means clustering algorithm (Everitt, 1993.) is traditionally used in data analysis. This minimizes the quality index, which is defined as entire points belonging to the cluster area, the distance to the cluster center (metric) (Agrawal et al., 1993). The metric in this context is the distance between the points included in the cluster (Li et al., 2004). Typically, in clustering algorithms the input data vector is compared to others or to predefined cluster center. The distance metric

also determines belonging to one or another cluster, thus determining the regularities in the multidimensional data samples, by attributing the input data to this or another class or cluster (Vitanyi, 2005).

Euclidean distance is the most widely used distance in clustering. This is the distance between the two-point coordinates in the multidimensional space corresponding to the length of the connecting segment, calculated from the formula (Agrawal et al., 1993):

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \quad (1)$$

Traditionally, in clustering algorithms the Euclidean distance is used, but choosing another metric is also a matter of discussion in some cases. It depends on the task being solved, the amount and complexity of the data.

In this research the similarity of objects is defined by the Euclidean distance: the smaller distance between two objects is, the more similar they are. The algorithm works in this way. At the beginning, the m centres c_j are set to some initial data points. If the training data is not ordered in a proper way, the first m training data is usually chosen as the initial set of function centres. Otherwise, m data points would be selected randomly. At step 2, each of the training patterns is assigned to the closest centre. At step 3, the centres are adjusted by taking the arithmetic average in each cluster group. Steps 2 and 3 will be repeated until each training pattern stays in its group, i.e., no reassignment of any pattern to a different group or previous group (see Table 1).

Table 1 ***K-Means clustering procedure***

Step 1. Initialize the function centres
Set the initial function centres to the first m training data or to the m randomly chosen training data.
Step 2. Group all patterns with the closet function centre
For each pattern xi, assign x_i to group j*, where $\left\| x_i - c_{j^*} \right\| = \min_j \left\| x_i - c_j \right\|$
Step 3. Compute the sample mean for the function centre
For each group c_j, $c_j = \frac{1}{m_j} \sum_{x_i \in group\ j} x_i$
where m_j is the number of patterns in group j.
Step 4. Repeat by going to step 2, until no change in cluster assignments

The operation of the algorithm results in the establishment of final cluster centers w_j, provided that the sum of square distances between all the points belonging to group j and the cluster center must be minimal.

An essential question in K-Means algorithm implementation is the determination of the number of clusters and initial centers. The simplest tasks assume that the number of clusters is known apriori and it is proposed to take the first m points of the training set for the initial values of the m cluster centers.

As an advantage of K-Means algorithm can be considered its popularity, high efficiency and simplicity of the procedure. But if the layout of the objects is heterogeneous, the algorithm may not produce good results. Then you need to change the parameters (number of clusters) and try again to repeat the algorithm's operations. The disadvantage is that the algorithm is not universal.

An example of the use of a clustering method for training purposes

To demonstrate the operation of a clustering algorithm, assume that we have 14 input vectors, which are split into two clusters. Using the K-Means clustering algorithm, it is necessary to determine the points and cluster centres belonging to each cluster (see Table 2).

Each input vector (or point) has two components: $x1$ un $x2$. The distribution of points in the 2-D plane is shown in Figure 2.

Table 2 ***Experimental data points***

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X1	1	3	6	10	2	2	5	6	4	8	8	4	9	1
X2	3	4	1	6	3	8	5	5	3	6	3	9	1	6

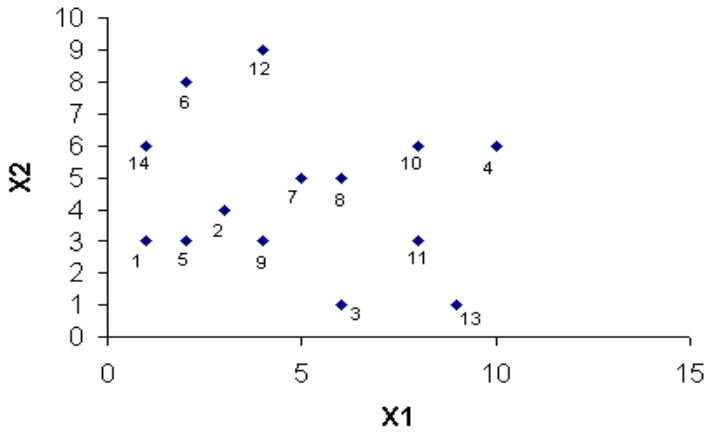


Figure 2 ***Initial data distribution***

In order to start using the clustering algorithm, it is necessary to determine the number of clusters and their initial centers. In this exercise we assume that input points are divided into two classes, so we will use two clusters.

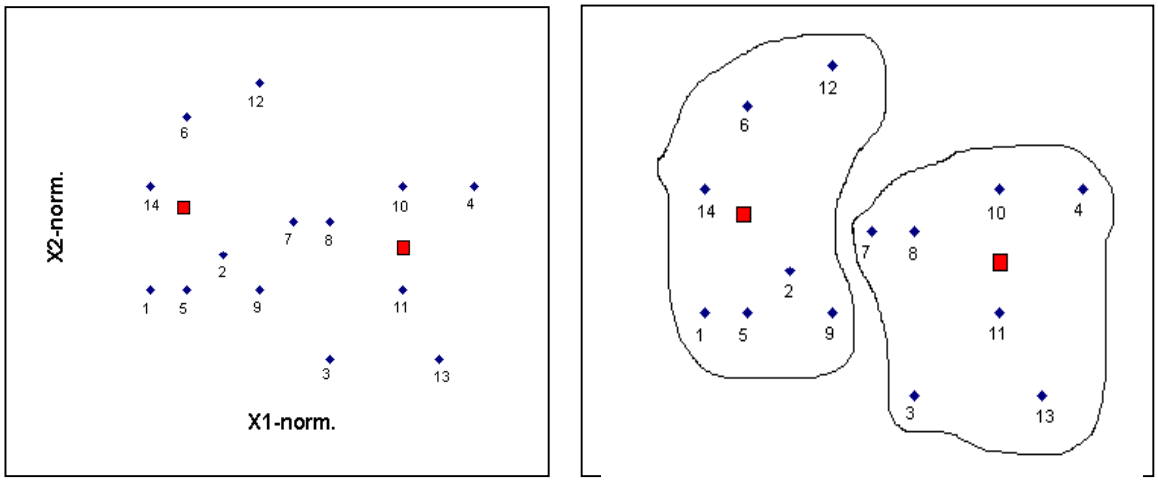


Figure 3 ***a) The distribution of points with an arbitrarily selected center; b) Separation of clusters after 1st iteration***

Figure 3a) below shows the distribution of points and the coordinate axes harvested for the sake of visibility. We approximately set the initial cluster centers with the arbitrarily selected coordinates. In the drawing, they are shown as squares. We start using the K-Means algorithm. Figure 3b) lists the points belonging to clusters after the first iteration:

Again, we calculate the average values for each cluster i.e. figure out new cluster centers. Since they are different from our arbitrarily selected initial cluster center, then we continue to apply clustering algorithm. The results are shown in Fig.4.

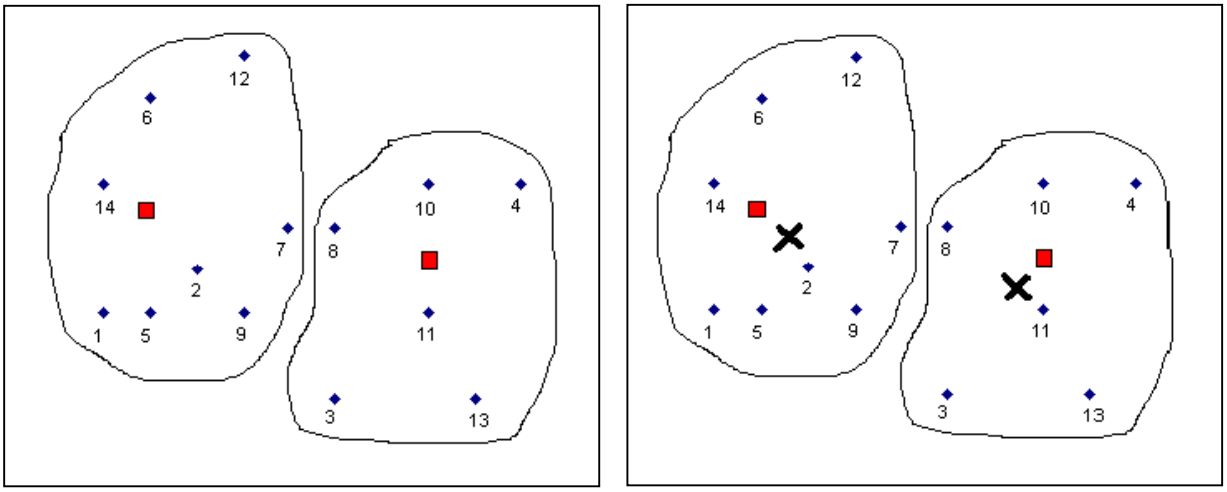


Figure 4 ***a) Separation of clusters after 2nd iteration. b) Separation of clusters after 3rd iteration***

As can be seen, after the 2nd iteration, the 7th data point has changed its class affiliation. Since the new cluster centers are different from the cluster centers obtained in the first iterations, then we continue to apply clustering algorithm.

In the third iteration the points have not changed their belonging to the clusters, i.e. cluster centers calculated in the second iteration remain unchanged. Consequently, it can be concluded that the use of the clustering algorithm in this case has set up cluster centers and the corresponding points from the trainee set are clustered. In Figure 4b, the new cluster centers are marked out with a cross.

Thus, with the help of clustering algorithm it was calculated that 8 data points: 1, 2, 5, 6, 7, 9, 12, 14 relate to cluster 1 and 6 data points: 3, 4, 8, 10, 11, 13 relate to cluster 2. Data is clustered.

Solution with SPSS and SPSS Modeler

The IBM SPSS software platform offers advanced statistical analysis, a vast library of machine-learning algorithms, text analysis, open-source extensibility, integration with big data and seamless deployment into applications. Its ease of use; flexibility and scalability make IBM SPSS accessible to users with all skill levels and outfits projects of all sizes and complexity to help you and your organization find new opportunities, improve efficiency and minimize risk. SPSS Statistics is leading statistical software used to solve a variety of business and research problems. It provides a range of techniques including ad-hoc analysis, hypothesis testing and reporting – making it easier to manage data, select and perform analyses (IBM Statistics, 2018).

Also with a help of SPSS Statistics package similar results are obtained in this case (see Table 3):

Table 3 ***SPSS Statistics results***

Number of Cases in each cluster	
Cluster 1	8
Cluster 2	6
Valid	14
Missing	0

It can be concluded that the results obtained by SPSS Statistics correspond to the manually calculated results of the clustering algorithm (8 data points relate to the cluster 1 and 6 data points - to the cluster 2).

The clustering results were tested with another IBM SPSS tool - IBM SPSS Modeler. SPSS Modeler is a leading visual data science and machine-learning solution. It helps enterprises accelerate time to value and achieve desired outcomes by speeding up operational tasks for data scientists. Leading

organizations worldwide rely on IBM for data preparation and discovery, predictive analytics, model management and deployment, and machine learning to monetize data assets. SPSS Modeler empowers organizations to tap into data assets and modern applications, with complete algorithms and models that are ready for immediate use (IBM Modeler, 2018).

The following clustering model for the implementation of the K-Means algorithm for the given data was performed with IBM SPSS Modeler (see Fig. 5).

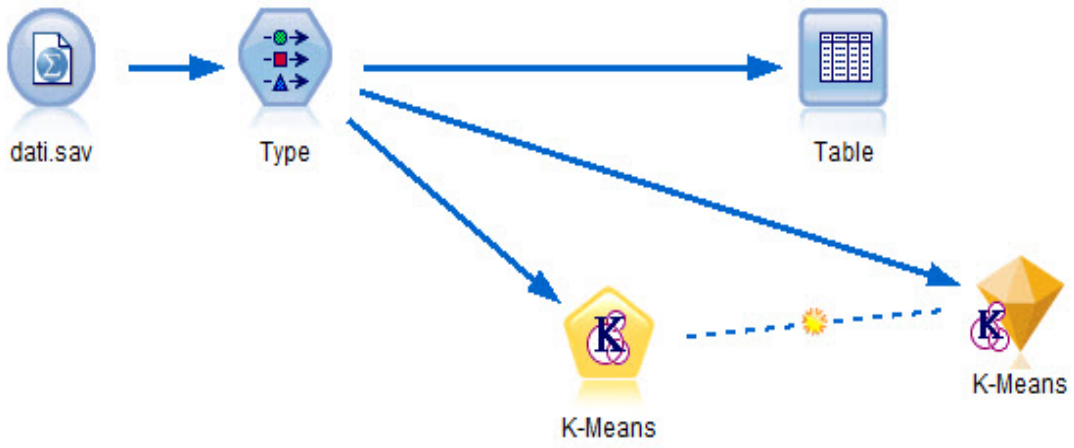
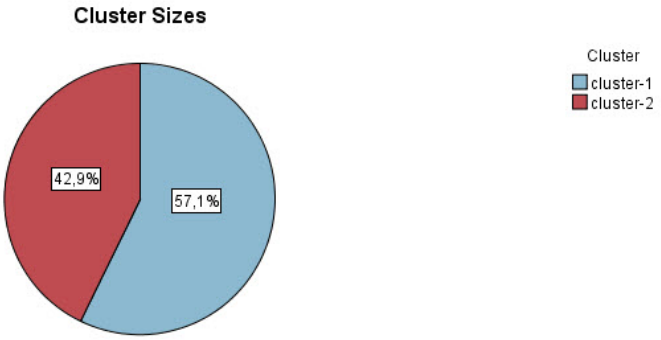


Figure 5 ***The clustering model in SPSS Modeler environment***

The sizes of the clusters are shown in Fig.6 and cluster centers in Fig.7. Similar results are obtained with SPSS Statistics - 6 data points are created in one cluster, 8 data points - in the second cluster.



Size of Smallest Cluster	6 (42,9%)
Size of Largest Cluster	8 (57,1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1,33

Figure 6 ***Cluster sizes***

Clusters

Input (Predictor) Importance
1,0 0,8 0,6 0,4 0,2 0,0

Cluster	cluster-1	cluster-2
Label		
Description		
Size	57,1% (8)	42,9% (6)
Inputs	X1 2,75	X1 7,8333
	X2 5,125	X2 3,6667

Figure 7 ***Cluster centers***

As a result, the following cluster centers C_1= (2,75; 5,125) and C_2= (7,833; 3,6667) are obtained corresponding to the manually calculated cluster centers.

It can be concluded that different methods give similar results according to the K-Means algorithm.

Conclusions

The purpose of the cluster analysis as one of the basic tasks of the intellectual data analysis is to search for independent groups and their characteristics in analytical data. Solving this problem allows for better understanding of the data, since clustering can be used in virtually any application area that requires experimental or statistical analysis of data.

All clustering algorithms have common parameters, the choice of which also characterizes clustering efficiency. The most important parameters characterizing clustering are: metrics (distance of cluster elements to the center), number of clusters k.

The paper provides a practical example that enables students to understand and begin to use the possibilities of modern Big Data analysis with the help of clustering algorithms.

Summary

The term "cluster analysis" dates back to 1939. It actually includes a complex of different classification algorithms. In the different fields of research, the live question is: "How to organize observable data in clearly viewed structures?" There is a view that, unlike many other statistical procedures, in most cases cluster analysis methods are used when there is no hypothesis about classes, but the data collection stage is still in progress. Methods of cluster analysis allow to divide the objects under investigation into groups with “similar” objects called clusters.

The cluster analysis process formally consists of the following steps:

- collecting of necessary data for analysis;
- determining the characteristic size and boundaries of class data (clusters);
- grouping of data in clusters;
- definition of class hierarchy and analysis of results.

The clustering algorithm K-Means minimizes the quality score, which is defined as a square sum of distance of all points belonging to the cluster area to the cluster centre. This procedure got its name because it is based on calculating the average distances of cluster groups to the cluster centers.

As a result of the algorithm, the final cluster centers are determined, provided that the sum of the squares of the distances between all the points belonging to the group and the cluster centers must be minimal.

As an advantage of the K-Means algorithm can be considered its popularity, high efficiency and simplicity of procedure. But if the placement of objects is heterogeneous, the algorithm may not achieve good results. Then it needs to change the parameters (number of clusters centers) and try to repeat the algorithm again. The drawback is that the algorithm is not universal.

An important issue in implementing the K-Means algorithm is determining the number of clusters and the initial centers. The simplest tasks assume that the number of clusters is known in advance. For the initial values of cluster centers it is suggested to take the first objects of the training cluster.

A possible solution would be to use the SPSS packages to implement various algorithms in Information Technology areas. Often, the analytical solution is much simpler than the visual SPSS model, but in perspective, for the sake of training it gives an understanding of the usefulness of using such models.

In the research part of the study the modelling capabilities in data mining studies were demonstrated, data clustering examples using IBM SPSS Statistics and Modeler were given.

The work provides a practical example that would enable students to understand and start mastering modern Big Data analysis capabilities with the help of clustering algorithms.

References

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. *Proc. 4th Int. Conf. On Foundations of Data Organizations and Algorithms*, Chicago, 69-84.

Aggarwal, C., & Reddy, C. (2013). *Data clustering: Algorithms and applications*. Chapman and Hall/CRC.

Everitt, B. (1993). *Cluster analysis*. Edward Arnold, London.

Gan, G. (2007). Data clustering: Theory, algorithms and applications. *ASA-SIAM series on Statistics and Applied Probability*, SIAM, Philadelphia, ASA, Alexandria, VA.

Han, J. (2001). *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.

IBM SPSS Statistics (2018). Retrieved from https://www.ibm.com/lv-en/marketplace/spss-statistics

IBM SPSS Modeler (2018). Retrieved from https://www.ibm.com/products/spss-modeler

Kaufman, L., & Rousseau, P. (2005). *Finding groups in data. An introduction to cluster analysis*. John Wiley & Sons.

Li, M. (2004). The similarity metric. *IEEE Transactions on Information Theory, 50*, 12, 3250-3264.

Vitanyi, P. (2005). *Universal similarity*. ITW2005, Rotorua, New Zealand.

Xu, R., & Wunch, D. (2009). *Clustering*. John Wiley & Sons.

Wierzchon, S., & Klopotek, M. (2018). *Modern Algorithms of Cluster Analysis*. Springer.