Journal of ELECTRICAL ENGINEERING, VOL. 64, NO. 2, 2013, 128–131

CLUSTER ANALYSIS — DATA MINING TECHNIQUE FOR DISCOVERING NATURAL GROUPINGS IN THE DATA

Elena Pastuchová [*] — **Štefánia Václavíková** [**]

Amount of data stored in databases has been growing rapidly. With the technology of pattern recognition and statistical and mathematical techniques sieved across the stored information, data mining helps researchers recognize important facts, relationships, trends, patterns, derogations and anomalies that might otherwise go undetected. One of the major data mining techniques is clustering In this paper some of clustering methods, helpful in many applications, are compared. We assess the suitability of the software that we used for clustering.

K e y w o r d s: data mining, clustering, K-means, self organizing maps, density based algorithm

1 INTRODUCTION

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields. The essence of cluster analysis is to identify clusters (groups) of objects such that the objects within a cluster are similar, while there is dissimilarity between the clusters. It means that objects in created cluster are as "close" as possible to one another and different groups as "far" as possible to one another, where distance is measured with respect to all available variables. In contrast to classification, that is process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible each from another, where distance is measured with respect to specific variable(s) you are trying to predict [5]. We focused on cluster analysis, because it has features that are very necessary in the examining the data. When there is lot of cases without visible aggregation, clustering algorithms can be used to find natural groupings. Clustering can also serve as useful data pre-processing step for identifying homogeneous groups on which to build models kept. There are many different algorithms available for performing cluster analysis .Also process of clustering is decision making process. After using different types of clustering on the same data, researchers usually obtain clusters that are not identical. These facts lead to necessity to measure the validity of clusters created by different methods. The problem is often in the selection of appropriate and available software. Our research was performed using statistical package programs PER SIMPLEX, MATLAB and R 2.12.0, which suitability we compare.

Described methodology could be found useful in technology practise for monitoring and diagnosis for instance of electrical equipments and systems allowing a simple identification of nominal regime taking into account admissible deviations as well as approaching of the monitored system to a'priori rated limit states.

2 CLUSTERING ALGORITHMS

Clustering methods that are most recently used in several applications are: K-means, density based algorithm and SOM.

2.1 K-mean algorithm

The algorithm allocates the data points (objects) into clusters, so as to minimize the sum of the squared distances between the data points and the center (mean) of the clusters. The centers of clusters are initialized by randomly selecting from the data. Then the data set is clustered in the process of assigning each point to the nearest center .When the data set has been identified, the average position of the data points within each cluster is calculated and the cluster center then moved to the average position. This process is repeated until all the cluster centers abide in a place.

2.2 Density based algorithm

Most partitioning methods cluster object based on the distance between objects. Such methods are able to find only spherically-shaped clusters and encounter difficulty in discovering clusters of arbitrary shapes. The development of DBSCAN has been based on the notion of density.The key idea is that for each point of a cluster the neighborhood of a given radius has to contain at least a

Slovak University of Technology in Bratislava, [*] Department of Mathematics, Institute of Informatics and Mathematics,Faculty of Electrical Engineering and Information Technology, elena.pastuchova@stuba.sk; [**] Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, stefania.vaclavikova@stuba.sk

DOI: 10.2478/jee-2013-0019, ISSN 1335-3632

minimum number of points, i.e. the density in the neighborhood has to exceed some threshold. Two global parameters for DBSCAN algorithms are:

1. ε – Maximum radius of the neighborhood,
2. MinPts – Minimum number of points in an ε-neighborhood of that point.

The *ε-neighborhood of a point* p, denoted by $Neps(p)$, is

$$Neps(p) = \{q \in D \mid \mathrm{dist}(p,q) \leq \varepsilon\}\,. \tag{1}$$

We say, that point p *is directly density-reachable from point* q, with respect to (*wrt*) some ε and *MinPts* if

$$p \in Neps(q) \text{ and } |Neps(q)| \leq MinPts\,, \tag{2}$$

where $|Neps(q)|$ is the number of elements of $Neps(q)$.

A point p is *density-reachable from a point* q, if there is a sequence of points $p_1, p_2, \ldots, p_n$, $p_1 = q$, $p_n = q$, such that p_{i+1} is directly density-reachable from p_i.

A point p is *density-connected to a point* q, if there is a point o such that both, p and q are density-reachable from o.

The algorithm of DBSCAN is as follows [2].

1. Select an arbitrary point p.
2. Retrieve all points density-reachable from p, *wrt* ε and *MinPts*.
3. If p is a core point, a cluster is formed.
4. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.

According to previous definitions, there are only three types of objects:

Core points – all points that have at least *MinPts* points in their ε-neighborhood;

Border points – all points that have less than *MinPts* points in their ε-neighborhood, but they are close enough to some core point;

Outliers – all other points.

2.3 Self organizing (Kohonen) maps (SOM)

The SOM algorithm is an unsupervised learning algorithm, where the learning algorithm establishes a topological relationship among input data. An attractive characteristic of the SOM is the ability to project high dimensional space (mainly spaces with more than 3 dimensions) into two or three dimensional spaces, that can be easily visualized [4]. Every map consists of some number of neurons (in most cases neurons are represented by hexagons). The aim of algorithm is to project all objects to neurons under the condition, that similar objects should be projected to similar neurons (this is called topology preserving). If we want to find clusters, we have to project all objects into two-dimensional space and identify similar objects. The advantage of this approach is that we also identify outliers in data that do not belong to any cluster. An algorithm that produces such a network can be briefly described as follows [3].

The SOM consists of two layers of neurons. The input layer is an array of M neurons. The neurons on the output layer are located on a grid with certain neighborhood relationship. The first step in constructing a SOM is to initialize the weight vectors.

$W_j = [w_{1j}, w_{2j}, \ldots, w_{Mj}]^\top$, $j = 1, 2, \ldots, N$ The SOM algorithm computes a similarity (distance) measure between the input vector $X = [x_1, x_2, \ldots, x_M]^\top$ and the weight vector $W_j = [w_{1j}, w_{2j}, \ldots, w_{Mj}]^\top$, $j = 1, 2, \ldots, N$ of each neuron u_j, the weight vector of each neuron has the same dimension as the input pattern. The Euclidean distance d_j between the weight vector W_j and input vector X is frequently used.

The output vector with the weight vector that is the smallest distance from the input vector is the winner. The weights of this winning neuron are adjusted in the direction of the input vector. Not only the winning neuron but also the neurons in the topological neighborhood of the winning neuron are affected by the competition. The winning neuron is the center of the topological neighborhood. The change to the weight vector W_j can be obtained as

$$\Delta W = \alpha h_j \big(X - W_j\big) \tag{3}$$

Where α is the learning-rate parameter of the algorithm and h_j is a topological neighborhood.

Hence, the updating weight vector $W_j(t+1)$ at time $t+1$ is defined

$$W_j(t+1) = W_j(t) + \alpha(t) h_j(t) \big(X - W_j(t)\big) \tag{4}$$

where $\alpha(t)$ and $h_j(t)$ are the learning-rate parameter and the topological neighborhood at time t. Applying (4) to all the neurons in the lattice that lie inside the topological neighborhood of winning neuron.

Upon repeated presentations of the training data, the weight vectors tend to move toward the input pattern due to the neighborhood updating. That is, the adjustment makes the weight vectors to be similar to the input pattern. The winning neuron shows the topological location of the input pattern. The neighborhood of the winning neuron shows the statistical distribution of the input pattern. The output of the SOM is obtained using a dynamic patterns grid, which shows a dynamic representation of the neurons that are winning each pattern.

3 EXPERIMENTAL RESULTS

The aim was to test various clustering techniques for pooling the analysed catchments of Slovakia into pooling groups based on seasonality of low flows lower than Q_{95}. In the beginning we realized cluster analysis using software Per Simplex, that was offered to Slovak Technical University for testing. This software create clusters like K-mean algorithm outlined below and then modified, if

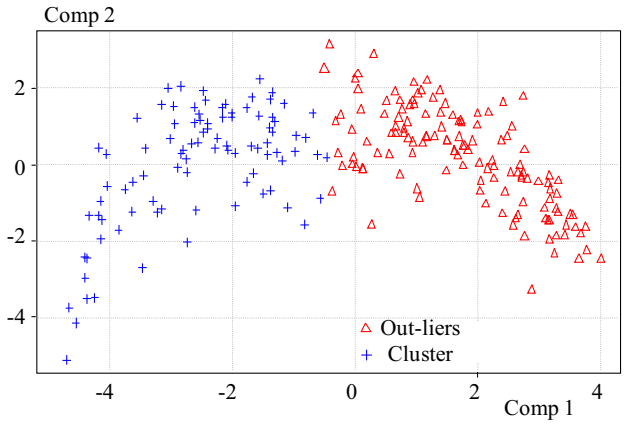

Fig. 1. Results of clustering based on K-means method

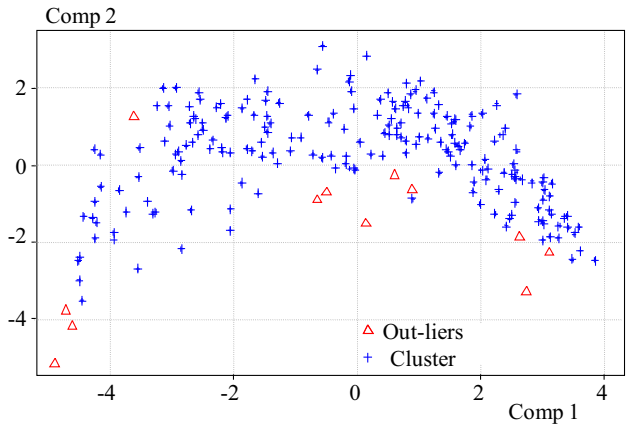

Fig. 2. Results of clustering by DBSCAN

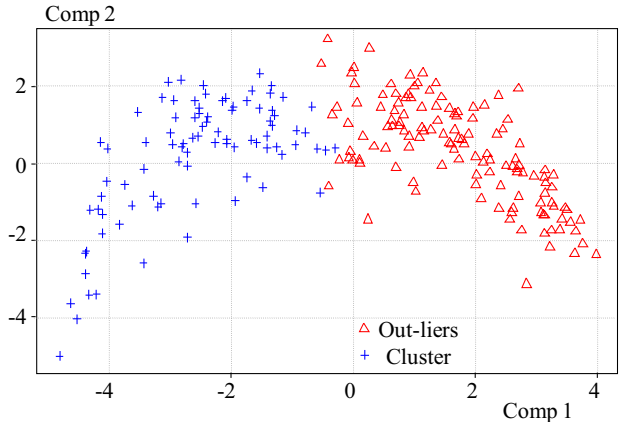

Fig. 3. Results of clustering using SOM

necessary, to improve the homogeneity of the groups as well as the size of the group.

Using K-means clustering, six clusters of homogeneous areas have been identified from the sample of 209 catchments where the basins within a cluster posses similar physiographic characteristics, while there is dissimilarity between the catchment in different clusters. The software offers clustering using K-means and "fuzzy" clustering, without getting to know the algorithm with which it operates [6]. Consequently, the following clustering method by SOM has been used Matlab 7.6.0.toolbox [9], while actual preprocessing, data clustering by K-means, DBSCAN, calculation silhouettes, Dunn coefficient was implemented in open-source R 2.12.0 program [8].

This analysis was conducted in a sample of 212 small and medium-sized basins from all over Slovakia. As a pooling variable, the relative frequency of occurrence of low flows, lower then Q_{95} was used.The application of the K-means method was analyzed by two clusters as the optimal solution (average silhouette coefficient was used to determine optimal number of clusters). The first cluster has 130 objects and covers most all regions of the Slovakia, while the second one has 82 objects, situated in the northern part of the country (see Fig. 1. where results from k-means algorithm are displayed in the first two principal components) [6].

The results of clustering with DBSCAN were significantly different, only one cluster was found and some objects were identified as outliers (see Fig. 2.). DBSCAN seems inappropriate for this dates, because two clusters are slightly overlapping and there is no clear separator between them which would have significantly low density of points (therefore only one large cluster of objects has been found.

In order to identify clusters using SOM algorithm, at the beginning, we used the so-called U-matrix and identified 2 clusters. PCA projection using first two principal components with SOM results can be seen in Fig. 3.

4 CONCLUSION

In this article three of clustering methods applied to identify the homogeneous areas are compared: K-means, DBSCAN and SOM.

The main advantage of K-means algorithm is its simplicity and speed which allows it to run on large datasets. K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular. An important question that needs to be answered before applying the K-means clustering algorithm is how many clusters are in the data. Determining the number of clusters is not trivial (we used average silhouette coefficient for determining true number of clusters).

DBSCAN is robust enough to identify clusters in noisy data, requires just two parameters and is mostly insensitive to the ordering of the points in the database. This algorithm is efficient even for large spatial databases, discovers clusters of arbitrary shape and does not need to know the number of clusters in the data a priori, as opposed to K-means. DBSCAN cannot cluster properly datasets with large differences in densities and therefore for this kind of data is unacceptable.

The SOM algorithm is very efficient in handling large datasets, robust even when the data set is noisy and "topology-preserving" feature superior to K-means methods. One of the advantages of SOM is clustering the data without specifying the number of clusters in advance.

While K-means and SOM give very similar results, DBSCAN has proved inappropriate for these data, because this method cannot cluster data with overlapping clusters.

Since program PERSIMPLEX was unusable for further analysis (for financial reasons), we used MATLAB and R 2.12.0 for further clustering.

The MATLAB Toolbox for SOM contains functions for creation, visualization and analysis of Self-Organizing Maps. The basic package can be used to preprocess data, initialize and train SOMs using a range of different kinds of topologies, visualize SOMs in various ways, and analyze the properties of the SOMs and data. With data mining in mind, SOM Toolbox is generally most suitable for data understanding and exploration, even though it may also be used for classification and modeling [10].

MATLAB and R have many common features. Contrary to R, MATLAB has many benefits, it is "easier" to use, one can use it without doing any programming, it is also a little faster with the normal configuration. MATLAB has lots of toolboxes for engineering applications. Overleaf R has statistical functionality difficult to find elsewhere and lots of statisticians use it. There is a very strong community around R, it is rapidly developing and free.

Using MATLAB and R, we got the same results for the SOM. Among all employed clustering methods, the results show that the SOM can determine the cluster membership more accurately than other methods. Procedures as well as the results indicate that the SOM is more quick and efficient than other methods.

Acknowledgement

This research was supported by grant 1/0103/10 of the Slovak Scientific Grant Agency.

References

[1] BURN, D. H.: Cluster Analysis as Applied to Regional Flood Frequency, Journal of Water Resources Planning and Management **115** (1989), 567–582.

[2] ESTER, M.—KRIEGEL, H.—SANDER, J.—XU, X.: A density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noiseinbook Proceedings of 2nd International Conference of Knowledge Discovery and Data Mining, Portland, OR.

[3] CHEN, L.—LIN, G. F.: Identification of Homogeneous Regions for Regional Frequency Analysis using the Self-Organizing Mapjour Journal of Hydrology.

[4] KOHONEN, T.: Self-Organizing Maps, the third, extended edition, Springer, 2001.

[5] LEON, A.: Enterprise Resource Planner, the third edition, Tata Mc Graw-Hill Publishig Company, New Delhi, 2008.

[6] PASTUCHOVÁ, E.—VÁCLAVÍKOVÁ, Š.: Cluster Analysis and its Application in Hydrology, Forum Statisticum No. 3 (2009), 135–139.

[7] PASTUCHOVÁ, E.—SABO, M.-KOHNOVÁ, S.: Comparison of Clustering Methods Applied to Hydrological Data, Forum Statisticum No. 7 (2011), 169–175.

[8] R Development Core Team 2011: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[9] SABO, M.—ŠTEVKOVÁ, A.—KOHNOVÁ, S.: Využitie samooganizujúcich sa máp pri regionálnej typizácii minimálnych prietokov na Slovensku, Acta Hydrologica Slovaca No. 1 (2011), 92–101.

[10] VENSANTO, J.—ALHONIEMI, E.—HIMBERG, J.—KIVILUOTO, K.—PARVIAINEN, J.: Self-Organizing Map for Data Mining in MATLAB: the SOM Toolbox, Simulation News Europe, 1999.

Received 7 September 2012

Elena Pastuchová, in 1978 graduated from the Faculty of Mathematics and Physics of the Comenius University, received her RNDr in 1990 and PhD degree in 2004. She is working at the Department of Mathematics of Institute of Computer Science and Mathematics of the Fakulty of Electrical Engineering and Information Technology, in Bratislava. Since 1990 she is involved in applied mathematics.

Štefánia Václavíková was born in 1970 in Slovakia, she was graduated from the Faculty of Mathematics and Physics of the Comenius University in 1993. She is working at the Department of Mathematics and Descriptive Geometry of The Faculty of Civil Engineering STU in Bratislava and she deals with applied mathematics.