cluster analysis, data mining,
k-medoids, BUPA

Jacek KOCYBA[1], Tomasz JACH[2], Agnieszka NOWAK-BRZEZIŃSKA[3]

MULTIDIMENSIONAL CLUSTERING DATA VISUALIZATION USING k-MEDOIDS ALGORITHM

The article presents the possibilities of using clustering algorithms to group and visualize data from blood tests of various people in the context of alcohol consumption impact on measured blood parameters. The presented results should be considered as the preliminary to the future works involving automatic visualization of medical data by using clustering algorithms. The authors present the results of clustering of the above data using k-medoids algorithm along with the proposition of visualization. The authors used as a set of input data "BUPA liver disorders" medical base taken from the Machine Learning Repository [7].

1. INTRODUCTION

Nowadays, information technology, and more particularly data analysis systems are used in most areas of science. Due to the enormous amount of data collected by humanity and all kinds of software, it is impossible to analyse it without using the appropriate computer applications [19]. It is quite often, when even the domain experts can not draw conclusions from raw data, whether due to its low readability caused by its complex characteristics, or because of enormous amount of the data. The purpose of the raw data analysis is to transform it into information, to get its unique and most useful features and to create a kind of summary. Standard statistical analysis is in many cases the simplest and yet effective solution, but only the use of methods, such as clustering algorithms, allows to obtain more valuable results of the process. In other words: to discover previously unknown dependences, and non-trivial conclusions [9], [20]. Cluster analysis as one of the data mining techniques becomes progressively more and more popular and there is a continuous development of the algorithms used in the process [10], [16]. It allows us to cluster data based on common characteristics, similarities in situations, when there are none pre-defined rules for doing so. These rules are being discovered during the execution of the clustering algorithm. The authors propose the use of the k-medoids algorithm for the analysis of medical data - blood test results in the context of the analysis of the measured parameters according to the amount of alcohol consumed by the subjects. One of the biggest problem using the k-medoids algorithm is to provide the correct visualization of the data, especially when there are more than two attributes involved in the clustering process. The authors propose an improvement in this area in order to facilitate the determination of cluster representatives. It is also crucial to show, that the data preparation process is vital to a proper clustering results. The experiments provided in this paper should be considered as the preliminary ones conducted to validate the usefulness of the k-medoids algorithm in different areas of data analysis.

As the authors browsed through the literature, they have not stumbled across many applications of k-medoids algorithm to medical data. The authors of [18] are presenting the faster method of computation

[1] kocybaj@tlen.pl.
[2] Institute of Computer Science, University of Silesia, Będzinska 39, 41–200 Sosnowiec, Poland, tomasz.jach@us.edu.pl.
[3] Institute of Computer Science, University of Silesia, Będzinska 39, 41–200 Sosnowiec, Poland, agnieszka.nowak@us.edu.pl.

of the k-means algorithm, but their visualization techniques are shown only for artificially created databases. The other works [5] also used the artificially generated data and thus provided no real-data visualisation.

Although, the k-medoids algorithm is quite popular among other authors [17], [4], [15], [21], [2], [6], [12], the proper visualisation of the results is still an open problem and therefore we propose our own preliminary method of visualisation. As the work in progress, the method is still in it's early stage.

2. THE STRUCTURE OF THE DATA

The analyzed dataset is taken from public archives of the Machine Learning Repository [7]. Its name is "BUPA liver disorders" and it contains blood tests of various people regularly drinking different amounts of alcohol. The file is in CSV (Comma Separated Values) format, and its structure is shown in Table 1.

Table 1. BUPA Database.

No	Attribute name	Meaning	Used in clustering?
1	MCV	Mean Corpuscular Value	YES
2	ALKPHOS	Alkaline Phosphatase (ALP)	YES
3	SGPT	Alanine Aminotransferase (ALAT, ALT)	YES
4	SGOT	Aspartate Aminotransferase (AspAT, AST)	YES
5	GAMMAGT	Gamma-glutamyl transpeptidase (GGTP, GGT)	YES
6	DRINKS	Number of half-pint equivalents of alcoholic beverages drunk per day	YES
7	SELECTOR	Field used to split data into two sets	NO

Dataset is complete - there are no missing values in any records. The SELECTOR attribute was omitted in the clustering process, because it does not contain any relevant information - its value (1 or 2) depends only on the DRINKS attribute, which was used in the clustering process.

3. DATA PREPROCESSING

All data is in the numerical form and thus can be used together with the k-medoids algorithm without extensive preprocessing. The only preprocessing elements that were used in this particular study was to eliminate the extreme values for each attribute of the data and perform the normalization of every attribute. This had to be done in order to make every attribute's influence the same. In order to implement the first stage of the preprocessing the interquartile range method [14] was used and the limit values were assumed as:

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ LowestLimit &= Q_1 - 1,5 \cdot IQR \\ HighestLimit &= Q_3 + 1,5 \cdot IQR \end{aligned}$$

4. DESCRIPTION OF THE SELECTED ALGORITHM

One of the most popular non-hierarchical clustering algorithm is k-medoids [13]. The principle of its operation is an extension of the k-means algorithm. Also in this case, the number of clusters k is determined before the first iteration and is constant during the execution of the algorithm. However, the algorithm does not calculate centroids like k-means, but medoids (m_c) instead [13], [22], which are considered the representative objects for each cluster. Their average similarity to other objects in

the same cluster is the highest. The main feature that distinguishes a medoid from a centroid is that a medoid is always one of the objects belonging to the cluster.

After selecting k random objects as the initial medoids, the remaining objects are assigned to the appropriate clusters by calculating the minimum distance according to a pre-selected measure. Subsequently the total cost TD (compactness of the clustering) is calculated [22]. Then, an attempt is made to switch initial medoids with objects not selected as medoids to improve the efficiency of the clustering. If the new TD is smaller in value, then re-grouping with the new medoids is done. This step is repeated until the stop condition (usually no change in medoids between two following iterations). Step-by-step example of this algorithm is presented in [11].

In this article the authors used two popular distance measures: Euclidean distance and Manhattan distance. Both of them are well known from literature [1], [8].

5. THE COMPUTATIONAL EXPERIMENTS

All the experiments have been conducted in a manner designed to detect the best parameters for the clustering data contained in the BUPA dataset. For this purpose specially written software was used. Every experiment was made using two attributes for clustering: Drinks and one of the others. For each pair of attributes three test were made: with no preprocessing of the data, only with normalization, with full preprocessing (outliers elimination and normalization).

As mentioned before, the data was preprocessed to eliminate outliers for each attribute taking part in clustering process using interquartile range method. Before this operation the dataset contained 345 objects and its statistics were as shown in Table 2.

Table 2. The structure of the attributes.

Attribute	Average	Median	Min	Max	1st quartile	3rd quartile	IQR
MCV	90,1594	90	65	103	87	93	6
ALP (ALKPHOS)	69,8696	67	23	138	57	80	23
ALT (SGPT)	30,4058	26	4	155	19	34	15
AST (SGOT)	24,6435	23	5	82	19	27	8
GGTP (GAMMAGT)	38,2841	25	5	297	15	46	31
DRINKS	3,4551	3	0	20	0,5	6	5,5

As the Table 2 shows, the attributes are only numerical. Some of theme have outliers which are easy to discover, while the other values of attributes are quite compact. After eliminating the outliers, 67 objects were omitted, so finally in clustering process 278 objects were considered.

Authors put the main focus on correct visualization of the data. Each time the best results have been achieved with full preprocessing and with use of the Euclidean measure. For this purpose, the authors present results of the clustering process on two dimensional charts (scatter-plots), thus each time visualization is done on two attributes of the dataset. The software allows creating charts using attributes independent from those used in the clustering process. However, changing the attributes used for visualization does not affect the results of the clustering process as it will be shown further.

After each experiment the quality of the obtained clusters was assessed by computing compactness of the clustering TDC [22]:

$$TD(C_i) = \sum_{p \in C_i} dist(p, m_{c_i}) \tag{1}$$

$$TD = \sum_{i=1}^{k} TD(C_i) \tag{2}$$

where:

- $TD(C_i)$ – compactness of single cluster
- TD – compactness of the whole clustering
- C_i – $i-th$ cluster

- p – object belonging to cluster
- m_{c_i}- medoid of C_i
- $dist(p, m_{c_i})$ – distance between object p and medoid m_{c_i}
- k - the number of clusters

Next assessment is performed by measuring how consistent are the generated clusters and how distinguished from each other are they.

Additionally authors analyse objects belonging to the clusters to designate representatives of the groups. This approach allowed to achieve very good results of the clustering process for data from Apache server logs [11] and the authors anticipate, that this approach would also lead to future improvement of the visualization of the medical data. After fine-tuning the algorithm, the representatives of the groups were found. This has led to improvement in server configuration and to optimization of website structure.

Below the authors presented some of the results of experiments with their proposition of visualization based on medical data.

5.1. IMPROVEMENT OF THE VISUALIZATION ACCORDING TO DATA PREPROCESSING

Fig. 1. Visualization of the clustering process without the use of data preprocessing.

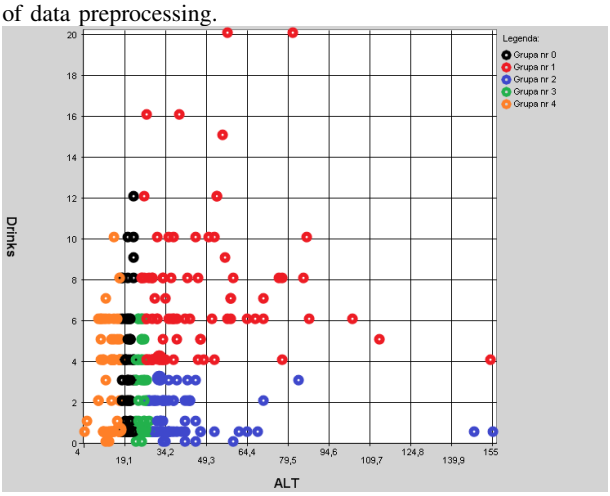
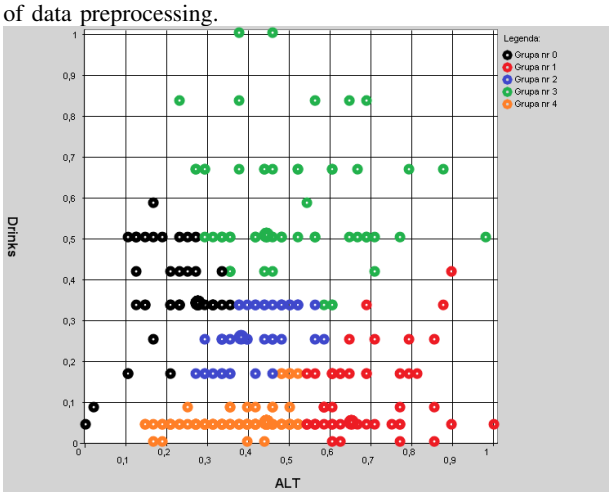


Fig. 2. Visualization of the clustering process with the use of data preprocessing.



As we can clearly see from comparison shown in Figures 1 and 2, preprocessing of the data gives a big improvement in readability of the visualization of the clustering process. Also eliminating outliers and normalizing the data ensures, that any attribute will not have a dominant influence on the clustering results.

5.2. VISUALIZATION FOR DIFFERENT DISTANCE MEASURES

In this experiment authors made a comparison between results of the process calculated with two distance measures: Manhattan and Euclidean. To precisely compare visualizations and cluster compactness, clustering process was executed each time with the same starting medoids.

In Figures 3 and 5-2 we can see the results of the experiments. Groups are easily distinguished on both charts. In this case there are no big differences, but for Euclidean measure TD is approx. 17% better ($TD_{Euc} = 37,031$ against $TD_{Man} = 44,569$).

Of course, there is a slight difference in numbers of members in groups:

- Manhattan: C[0]=81, C[1]=28, C[2]=39, C[3]=51, C[4]=79
- Euclidean: C[0]=80, C[1]=26, C[2]=41, C[3]=52, C[4]=79

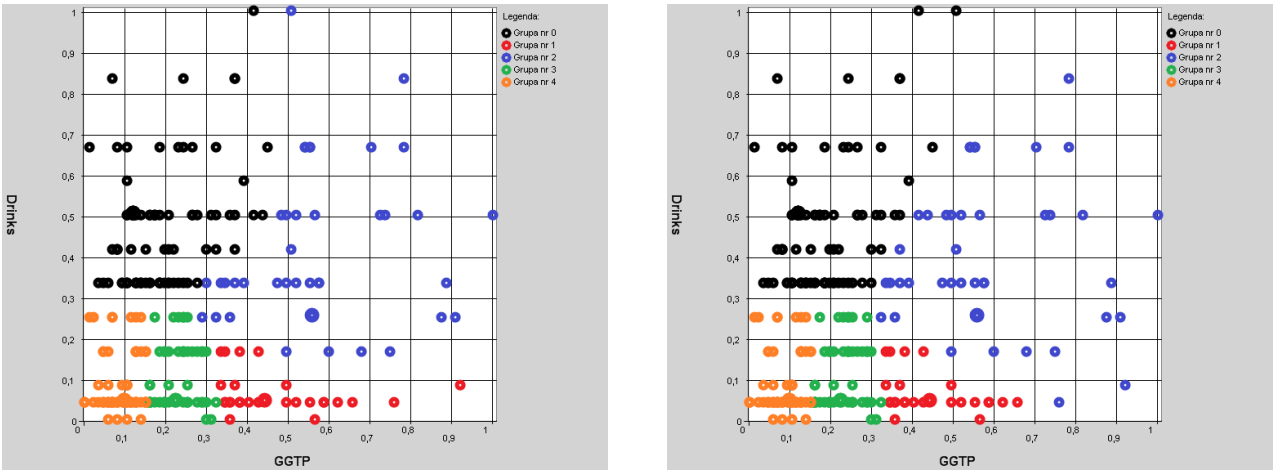


Fig. 3. Results obtained using Manhattan distance measure.

5.3. VISUALIZATION FOR DIFFERENT ATTRIBUTES

Below authors presents results of clustering using different attributes of the data (every time Drinks and one of the others). The k parameter (number of groups) is constant and has value of 5. The value of this parameter was chosen after performing many experiments on this particular dataset.

On charts presented on page 68 the shapes of groups and number of members in each groups is quite different. Every time we can see clearly distinguished groups. The best compactness of whole grouping ($TD = 35,387$) was achieved using attributes: Drinks and GGTP, the worst ($TD = 54,971$) with attributes: Drinks and ALP, what gives over 35% of difference.

According to conducted experiments, we can see, that after simple statistical analysis, grouping results for attributes: Drinks and GGTP gives clear sight on representatives of the groups. For example:

- - group number 0 (black dots on the visualization) has close to each other average values (after normalization) of the MCV, ALP, ALT, AST parameters and low average value of GGTP parameter. Average value of Drinks parameter is very low. Standard deviation of the first four parameters is close to 0,16 and close to 0,04 for GGTP and Drinks. The authors present some statistics of this group (values after normalization) in table 3 on page 67:

Table 3. Statistics of the first selected group (averages and standard deviations).

$AVG(MCV) = 0,469\ (89,32)$	$\sigma(MCV) = 0,15$
$AVG(ALP) = 0,512\ (67,01)$	$\sigma(ALP) = 0,17$
$AVG(ALT) = 0,378\ (22,13)$	$\sigma(ALT) = 0,155$
$AVG(AST) = 0,416\ (20,07)$	$\sigma(AST) = 0,185$
$AVG(GGTP) = 0,111\ (14,68)$	$\sigma(GGTP) = 0,038$
$AVG(Drinks) = 0,065\ (0,78)$	$\sigma(Drinks) = 0,045$

Compactness of the group:

$$TDC[0] = 3,654$$

Number of members:

$$C[0] = 72$$

As we can see, TDC value for this group is very low, what means very good quality of clustering, especially in correlation with the number of members of this group. The representative of this group is a person with quite healthy liver (all tests in norm) who drinks less than one drink a day.

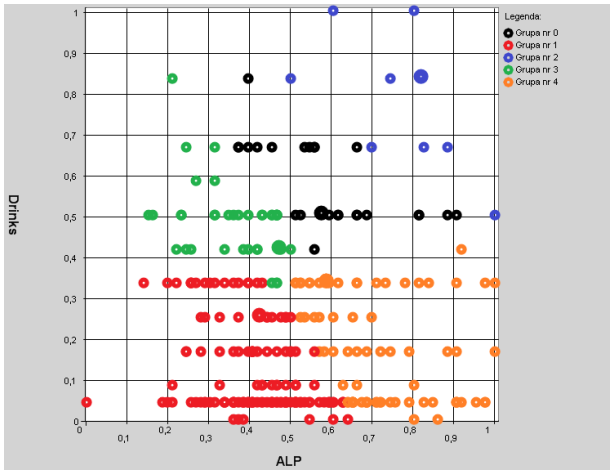


Fig. 4. Results for clustering using Big Drinks and ALP attributes.

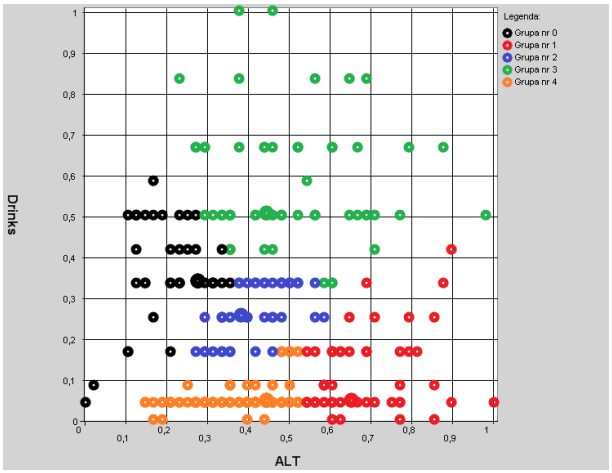


Fig. 5. Results for clustering using Drinks and SGPT attributes.

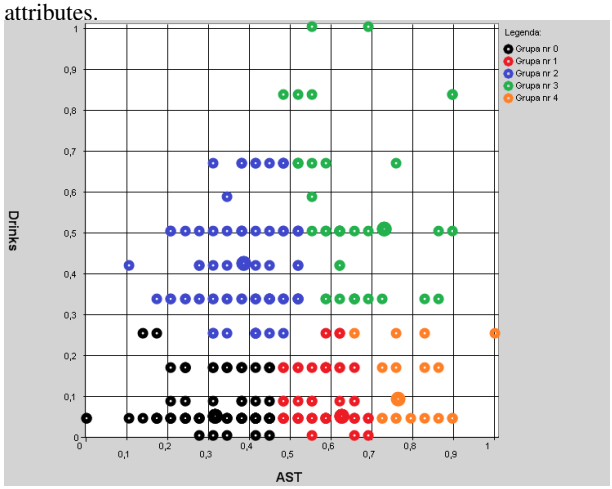


Fig. 6. Results for clustering using Drinks and SGOT attributes.

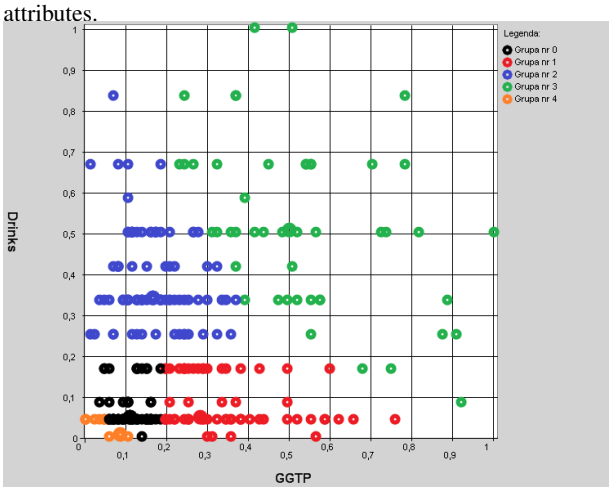


Fig. 7. Results for clustering using Drinks and GAMMAGT attributes.

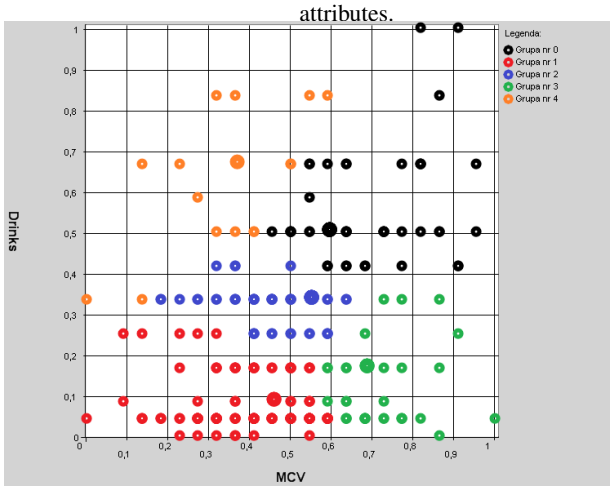


Fig. 8. Results for clustering using Drinks and MCV attributes.

- - group number 3 (green dots on the visualization) has an average values of each attribute almost at the same level (values after normalization), also standard deviation for all of them is on the same level (0,2). The authors present some statistics of this group (values after normalization) in table 4 on page 69:

Table 4. Statistics of the second selected group (averages and standard deviations).

$AVG(MCV) = 0,598\ (92,15)$	$\sigma(MCV) = 0,221$
$AVG(ALP) = 0,543\ (69,69)$	$\sigma(ALP) = 0,213$
$AVG(ALT) = 0,533\ (29,58)$	$\sigma(ALT) = 0,19$
$AVG(AST) = 0,568\ (24,46)$	$\sigma(AST) = 0,178$
$AVG(GGTP) = 0,537\ (51,75)$	$\sigma(GGTP) = 0,196$
$AVG(Drinks) = 0,519\ (6,23)$	$\sigma(Drinks) = 0,207$

Compactness of the group:

$$TDC[3] = 11,975$$

Number of members:

$$C[3] = 48$$

Objects in this group have the highest average value of Drinks attribute. From a medical point of view representative of this group is a man, who drinks a little more than 6 drinks a day, MCV is just under upper limit of the norm, ALP, ALT, AST are also in norm, but GGTP is 35% over upper limit of the norm which is 38 for males. In conjunction with relatively high alcohol consumption this may indicate liver disease, but the correct diagnosis would require additional tests.

Because of the k-medoids algorithm is sensitive to starting conditions, this results can vary in other tries. The authors wish to improve this inconvenience in their further research e.g. by choosing the most dissimilar initial medoids.

6. CONCLUSIONS

From all of the above, authors drawn the conclusion, that the proposed approach is also possible to apply to multi-dimensional medical data. Proposed solution is not yet in it's final state. Currently authors are working on a proper and clearly legible visualization for multi-dimensional data (when there are three and more attributes involved in clustering and visualization). Also, a very important issue is the ability to automatically determine representatives for the clusters. This is the problem which authors plan to resolve in their further research.

BIBLIOGRAPHY

[1] ABONYI J., FEIL B., Cluster Analysis for Data Mining and System Identification, Birkhäuser Verlag AG, 2007.
[2] ALSABTI K., RANKA S., SINGH V., An Efficient k-means Clustering Algorithm, Proc. First Workshop High Performance Data Mining, 1998.
[3] CIOS K. J., PEDRYCZ W., ŚWINIARSKI R. W., KURGAN L.A., Data mining. A Knowledge Discovery Approach, Springer Science+Business Media.
[4] CHU S. C., RODDICK J. F., CHEN T. Y., PAN J. S., Efficient search approaches for k-medoids-based algorithms, TENCON '02, Proceedings, 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002.
[5] CHU S. C., RODDICK J. F., PAN J. S., An Efficient K -MedoidsBased Algorithm Using Previous Medoid Index, Triangular Inequality Elimination Criteria, and Partial Distance Search, 4th International Conference on Data Warehousing and Knowledge Discovery, 2002.
[6] ESTER M., KRIEGEL H. P., SANDER J., XU X., A density -based algorithm for discovering clusters in large spatial databases, Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'96), 1996.

[7] FRANK A., ASUNCION A., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.
[8] JAIN A. K., DUBES R. C., Algorithms for clustering data, New Jersey: Prentice Hall, 1988.
[9] KAUFMAN L., MASSART D. L., The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, 1983.
[10] KAUFMAN L., ROUSSEEUW P., Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley, 1990.
[11] KOCYBA J., Grupowanie danych zawartych w logach systemowych, Master Thesis, University of Silesia, 2013.
[12] MATHEUS C. J., CHAN P. K., PIATETSKY-SHAPIRO G., Systems for Knowledge Discovery in Databases, IEEE Transactions on Knowledge and Data Engineering, 1993.
[13] MERCER D. P., Clustering large datasets, Linacre College, 2003.
[14] MOORE D. S., The Basic Practice of Statistics, Purdue University, 2010.
[15] MUMTAZL K., DURAISWAMY K., A Novel Density based improved k-means Clustering Algorithm - kmeans International Journal on Computer Science and Engineering, 2010.
[16] MYATT G. J., Making Sense of Data A Practical Guide to Exploratory Data Analysis and Data Mining, New Jersey: John Wiley and Sons, Inc, 2007.
[17] NG R.T., HAN J., CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on knowledge and data engineering, 2002.
[18] PARK H. S., JUN C. H., A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 2009.
[19] SIMIŃSKI R., NOWAK-BRZEZIŃKA A., JACH T., XIĘSKI T., Towards a practical approach to discover internal dependencies in rule-based knowledge bases, Rough Sets and Knowledge Technology, Lecture Notes in Computer Science, Springer /Heidelberg, Berlin, 2011, pp. 232-237.
[20] SUH S. C., Practical Applications of Data Mining, 2012.
[21] ZHANG T., RAMAKRISHNAN R., LIVNY M., BIRCH: An efficient data clustering method for very large databases, In: SIGMOD Conference, 1996.
[22] ZHONG N., LIU J., YAO Y., Web Intelligence Meets Brain Informatics, 2006.