Dendrograms-based disclosure method for evaluating cluster analysis in the IoT domain

Roman Kaminskyy [1], Natalya Shakhovska[1], Natalia Kryvinska [2], Muhammad Younas [2]

[1] Artificial intelligence department, Institute of computer sciences and information technologies, Lviv Polytechnic National University, Lviv, Ukraine, nataliya.b.shakhovska@lpnu.ua
[2] Department of Information Systems, Comenius University in Bratislava, Bratislava, 81499, Slovakia, natalia.kryvinska@uniba.sk
[3] School of Engineering, Computing and Mathematics, Oxford Brookes University, Oxford, UK

Abstract: The Internet of Things (IoT) generates huge amount of data at an extremely fast pace. Thus, it is important to classify such data objects into different groups or clusters in order to gain some valuable insights from data. This paper aims to develop a dendrograms-based method for 3D visualization of hierarchical clustering for multidimensional data which can be collected from IoT devices and open databases. This method is built on hierarchical clustering algorithm which is simple and efficient. It presents areas of the selected clusters and their objects on a plane, according to the coordinates defined by the open dendrogram. It defines rules for visualization of the dendrogram and allows to find the nature of clusters. The paper also proposes quantitative indicators of localization of objects and evaluation of clusters being formed. The proposed method is evaluated using different datasets. The proposed method significantly improves the quality of visualization and evaluation of cluster analysis results. It is also efficient as the time complexity is significantly less for factorial analysis.

Keywords: clustering, dengrogdam disclourse, quality of clustering, 3D visualization, data mining, quality of grouping

1. Introduction

IoT is penetrating into every facet of modern lives ranging from smart cities to autonomous vehicles through to healthcare to smart utility services. For instance, Internet of Things for Industry (Industry 4.0), has become the de facto standard for European businesses. IoT generates a large volume of data through different means and devices such as sensors, wearables, and smart devices.

The increase in the volume of IoT data leads to the problem of data accumulation, cleaning, (pre)processing and analysis. One of the data preprocessing types is the cluster analysis which is an objective method of classification. It provides an appropriate choice of processing methods as well as the visualization and interpretation of the collected data, which are termed as multidimensional objects.

The outliers detection is one of the most important task in IoT solution. The destroyed sensors or problems with data transmissions can affect the automatic decision making procedure. That is why we must find not only the source of the problem, but analyze closest to these objects and visualize the result of analysis. The most valuable feature of cluster analysis is the representation of the result by an image of a dendrogram that reflects a particular hierarchy of relationships between the selected clusters and their objects. In recent years, cluster analysis has attracted significant attention from the research community. A large number of publications is devoted to cluster analysis such as scientific articles, textbooks and monographs, which report on different approaches to conducting and interpreting cluster analysis results.

This paper aims to develop a dendrograms-based method for 3D visualization of hierarchical clustering for multidimensional data which can be collected from IoT devices and open databases. The method is built on hierarchical clustering algorithm which is simple and efficient. It presents areas of the

selected clusters and their objects on a plane, according to the coordinates defined by the open dendrogram. It defines rules for visualization of the dendrogram and allows to find the nature of clusters. The paper also proposes quantitative indicators of localization of objects and evaluation of clusters being formed. The proposed method is evaluated using different datasets. The proposed method significantly improves the quality of visualization and evaluation of cluster analysis results. It is also efficient as the time complexity is significantly less for factorial analysis.

The paper is structured as follow. Section 2 reviews related work on hierarchical clustering methods. The dendrogram development techniques are analyzed. Section 3 presents the proposed method for disclosure of the dendrogram. Section 4 presents the experimentation and results. Section 5 concludes the paper.

2. Related works

A overview of cluster analysis publications is provided in [1]. This review pays particular attention to the optimal number of clusters. The strategy of agglomerative hierarchical cluster analysis is given in [2]. However, it falls short of discussing intersection of clusters. The algorithms of agglomerative hierarchical clustering and methods for determining similarity between clusters are analyzed in [3]. The objects are presented only on the plane; that is why intersection of clusters is possible too.

The work in [4] covers clustering algorithms, their applications and advantages. The time complexity for several clustering algorithms are estimated. For IoT-systems this is very important to reduce time and computation complexities. That is why the time complexity of proposed algorithm should be estimated and compared wth existing approaches.

Paper [5] presents the clustering algorithm using single-linkage metric, which provides the shortest trajectory for all object connections. Information technology of hierarchical agglomerative analysis of one-dimensional data presented by samples is given in [6, 7]. However, the comparison with another metrics is not discussed . This is very important for the cluster shape evaluation.

The original interpretation of the dendrogram is given in [8]. The clusters are represented on the plane, since only two features are used. In [9, 10] results of clustering of economic data and their display on the geographical map are presented. In [11] and [12], the hierarchical cluster analysis is compared with other cluster technics.

Different approaches to the interpretation of dendrograms, as well as variants of hierarchical agglomerative and divisive analysis are given in [13 – 14]. Dendrogram visualization techniques are presented in [15, 16]. The difference between metrics is presented. This method, in our opinion, is quite simple, has the ability to choose the appropriate strategy combining elements and for correct numerical data given by the proximity matrix, ensures their objective and accurate clustering.

For the streaming data as well as for multidimensional data, the one known feature is a representative sample size. For such objects, agglomerative hierarchical analysis is preferably used to subdivide such groups into subgroups [17] and descriptive statistics indicators are used as features. Moreover, the result is the construction of the dendrogram. For each cluster, it is also possible to determine the generalized (average) quantitative values of the features of its objects on which this cluster analysis was performed [11, 12].

In [18], the genetic algorithm monarch butterfly optimization (MBO) is presented. The population is divided into two subpopulations according to k-means clustering. Each generation is built using minimal intra-distance. The drawback of this algorithm is dependence of cluster parameters from number of clusters. The same feature of clustering process is for elephant herding optimization algorithm [19].

The outliers detection in IoT-solutions based on dendrogram with combination of Long-short Term Memory neural network (LSRM) is presented in [20]. Hierarchical clustering is used for the outlier detecting by finding correlated sensors. Authors in [21] propose to use clustering for Damage Detection Scheme. The

k-means is improved using normalized Euclidean Ichino-Yaguchi distance. However, both approach cannot be used for data visualization in 3D.

So, vast majority of cluster analysis methods are related to building dendrogram and visualizing the clusters. In addition, existing literature focuses on individual clusters, wherein objects are considered as homogeneous such that they are processed by similar methods.

3. Dendrogram: basic concepts and design

There exist different types of dendrograms for interpreting the result of classification of multidimensional objects. The diversity between the types of dendrograms significantly complicates the task of comparing results obtained from experimentation. The purpose of cluster analysis is to present result of a dendrogram, by which the number of clusters and objects belonging to each cluster are determined. Such classification is carried out by specific algorithms. In general, an algorithm uses metric and a specific strategy for combining objects into their respective clusters.

Cluster analysis generally assists in solving the following problems:

- identifying and understanding structure of data;
- establishing uniformity across the whole group of objects;
- detection of atypical objects;
- justify the number of clusters;
- ensure a high degree of similarity of objects within each cluster;
- representation of the individual objects that do not fit into any of the clusters.

Different clustering technics provide partition information about the structure of clusters, such as the homogeneity of objects in a cluster, the variability (density of objects), and the shape of the cluster. Taking into account, all such characteristics, is important for the analysis of different tasks.

The scale of the dendrogram determines the distance between the clusters, but their localization on the plane is difficult to predict.

A close look at the dendrogram, rotated downwards, may reveal an analogy with the shadow on the wall of a nude young tree illuminated by the rays of the sun. Therefore, we can assume that the constructed dendrogram also has a crown on the cross-sectional plane of which is the localization of objects. It is believed that presentation and interpretation of results of cluster analysis of the dendrogram can be considered in this sense. The dendrogram is actually the projection of a dual graph onto a plane. This plane is analogous to the Cartesian plane but has two different scales: metric for distances (ordinate) and nominal for objects and nodes (abscissa). It does not take into account that the root of the tree is at the top. However, it can be assumed that the tree itself is a volumetric figure in 3D space. Therefore, when we perceive the dendrogram, we cannot say which objects are closer to us and which ones are further. In other words, all the objects are same on the line in 2D representation.

When we present the dendrogram as a tree (root below), it means that the objects on this horizontal line lie in the plane of cross section of the tree. Therefore, the task of visualizing these clusters arises, and it is possible to determine certain quantitative characteristics for such clusters. In addition, it can be assumed that clusters may overlap with one other in this cross-sectional plane.

The above observation reveals that 3D visualization is required. This then leads to the definition of the following hypothesis.

Hypothesis. If a dendrogram is rotated at a root below and is viewed as a 3D tree, the object line will correspond to the cross-sectional plane of the branches.

When viewed from top of the dendrogram, the ends of the truncated branches will be distributed in some way in this plane. This will then enable some distribution of clusters, objects in clusters and also distribution of nodes on a plane. This plane can be called 3D cross section of the dendrogram along the

object line or the cluster plane. The distance between objects will be preserved so that objects are visually more precise and more clearer when they are viewed (represented) in clusters. This also helps in appropriately organizing the clusters. This representation is intuitive and dichotomously limited, since only the following pairs are used: object-object, object-group, group-group.

The purpose of the paper is to solve the problem of 3D presentation of dendrogram by exploiting the disclosing visual and informational content of the dendrogram. Consider a dendrogram with a root node at the top and on the horizontal line. This paper proposes to determine the position of the root node and its nearest left and right nodes in the dendrogram. The ordinate of the dendrogram determines the lengths of these distances. Consequently, three nodes on this horizontal line are built on the left, root and right of the dendrogram. Next, a vertical line through the left node will be built. The dimensions to this node are determined by distances from that left node to the nodes it covers. Moreover, the left nodes will be situated at the top of the vertical line, and the right - at the bottom. The same procedure is invoked with the right upper node, but the difference is in creation of the vertical line with right node above and left node below. These steps are continued for other nodes, following the hierarchy established for them.

In fact, in this way the plane of cross section of the dendrogram is formed and visualized. However, unlike conventional visualization, this approach is characterized by the fact that all branches will have a length value at which certain numerical characteristics of the clusters can be obtained. In other words, the analog of such construction is the construction of a recursive H-fractal, known as the "Mandelbrot tree". In this case, this representation of the cross section of the dendrogram is a very irregular H-fractal.

4. Development Methodology

This section describes the methodology for development of the dendrogram. The methodology comprises various phases, which are illustrated as follow.

First, dataset is described. Descriptive statistics based on hierarchical clustering results shows object distribution in the plane. Next, the rules for the nodes numbering in the agglomerative hierarchical cluster analysis are proposed. Construction of the dendrogram based on these rules can be used for density evaluation. The last step is disclosure of dendrogram. The result of this step is 3d-visualization and shape of the cluster evaluation. Perimeter and Square of created cluster can be calculated too.

4.1. Data collection

The proposed approach takes into account two datasets each having different features or characteristics. Our research is based on dataset of Centre for Advanced Studies in Adaptive Systems (CASAS) smart home project [22].

Two dataset are built based on this dataset. After preprocessing stage and data transformation the first dataset is developed. The characteristics of the first dataset are presented as indicators of descriptive statistics. We use eight features. The second dataset is built on expert evaluation of the initial dataset parameters. The scale of 10 points is used for this. After creating the object-property table and normalization of data, their values occur in the interval $0…1$.

Euclidean metric was used to calculate the table of similarities between the objects in both datasets. A hierarchical agglomerative cluster analysis was performed [17]. A flexible strategy was used as an aggregation strategy, which is applicable to any proximity measure.

The process was implemented in R programming language using RStudio. Results of cluster analysis are shown in Fig. 1. The results show that it is rather complex to find dependencies between clusters using traditional approach.

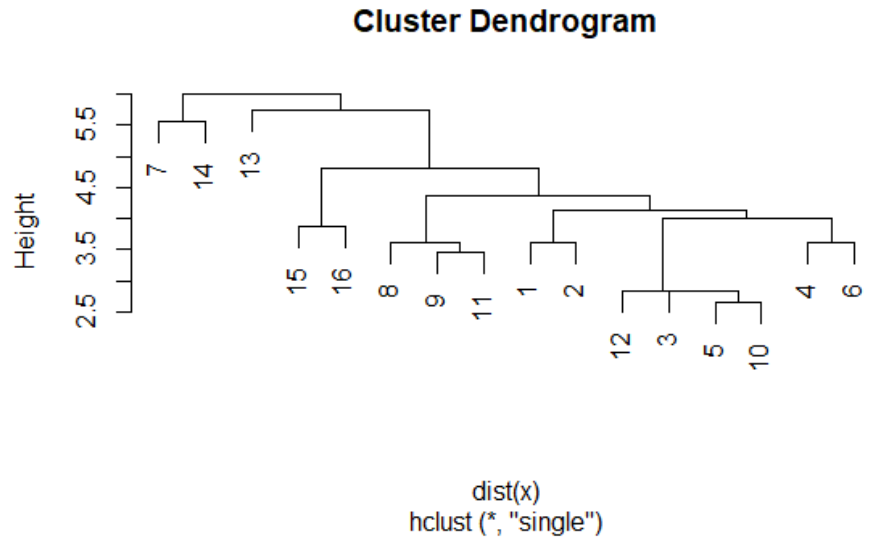

Fig. 1. Hierarchical clustering results using for the first dataset (descriptive statistics)

Conducting agglomerative hierarchical cluster analysis for both datasets ensured the parameter values of their dendrograms. The values of these parameters are distances between objects, object and group, and between groups. The results are shown in Table 1 and Table 2. The first column shows the numbering of groupings of objects of classification (into groups). The second column shows objects and groups that merge. The symbol "∪" means the union of two elements: an object with an object, an object with a group, a group with a group. Object letters are highlighted in bold. The third column shows metric values (distances) for constructing the dendrogram.

Table 1. Results and analysis of the first dataset

Nodes	Groups	Metrics
14	**2** ∪ 4	0,08
15	**8** ∪ **9**	0,09
16	**11** ∪ **13**	0,09
17	**1** ∪ **6**	0,14
18	**7** ∪ 16	0,15
19	**5** ∪ 15	0,16
20	**3** ∪ **12**	0,19
21	**10** ∪ 17	0,27
22	18 ∪ 19	0,34
23	14 ∪ 20	0,36
24	21 ∪ 23	0,60
25	22 ∪ 24	0,92

Table 2. Result and analysis of the second dataset

Nodes	Groups	Metrics
21	**3** ∪ **5**	6.00
22	**10** ∪ **12**	6.00
23	**1** ∪ **2**	7.00
24	**17** ∪ **18**	8.00
25	**4** ∪ 21	8.50
26	**8** ∪ **9**	9.00
27	**15** ∪ **16**	9.00
28	**11** ∪ 26	9.63
29	**20** ∪ 22	9.75
30	**19** ∪ 23	10.13
31	**14** ∪ 24	10.50
32	**6** ∪ 25	11.16
33	**7** ∪ **13**	13.00
34	29 ∪ 30	13.89
35	28 ∪ 32	15.43
36	27 ∪ 33	16.59
37	34 ∪ 35	22.51
38	31 ∪ 36	24.94
39	37 ∪ 38	54.67

The dendrograms are built based on the distance metric between objects, between an object and a group of already merged objects, and between the two groups.

3.2. Rules for construction of the dendrogram

The construction of a dendrogram in a cluster analysis may have different variants of branch distribution [15].

Hierarchical clustering allows an arbitrary position of objects on the abscissa line for dendrogram. As a result, for the same data, the dendrograms obtained by different researchers may appear visually different. Though correct, an arbitrary placement of objects on the line of dendrogram often complicates its visual perception. Existing programs build dendrograms differently. They often require manual adjustments and make them more comprehensible.

The numbering of nodes in the agglomerative hierarchical cluster analysis strictly corresponds to the levels of hierarchy of dendrogram. Such numbering is easily perceived if the sequence of node numbers is ordered. In addition, it is better to visually combine objects if an object with a smaller value is on the left and a larger one is on the right. If the object is merged with a group, then it should be added on the left. The following rules for the construction of dendrogram (as shown in Fig. 2) are defined.

Rule 0. The dendrogram building starts from its root node.

The node 25 is placed on top. This node integrates two nodes: node 22 and node 24.

Rule 1. Since mergers were done in pairs, nodes and objects with a smaller number are written on the left and, nodes and objects with a larger number are written on the right.

Node 22 integrates two nodes: 18 and 19. According to rule 1, node 18 will be located on the left and node 19 on the right. Node 24 also integrates two nodes: 21 and 23. According to rule 1 node 21 is placed on the left and node 23 is placed on the right.

Rule 2. The vertical distance between nodes must correspond to their metric given in the table on the scale of y-axis. This rule also determines the position of objects on the abscissa.

Node 18 integrates object 7 and group 16 created from objects 11 and 13. Node 19 integrates object 5 and group 15 created from objects 8 and 9. Node 21 integrates object 10 and group 17 that integrates objects 1 and 6. 8. Node 23 integrates two groups 14 and 20, with node 14 being the union of objects 2 and 4, and node 20 being the union of objects 3 and 12.

Schematic representation of the dendrogram based on in Table 1, is shown in Fig. 2.

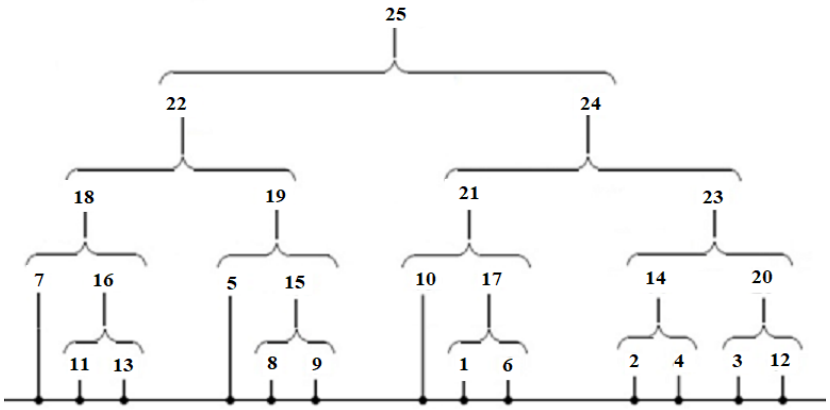

Fig. 2. Schematic representation of the structure of dendrogram (c.f. Table 1).

3.3. Cconstruction of the dendrogram

Fig. 3 shows the construction of dendrogram which is based on the rules defined in the preceding section.

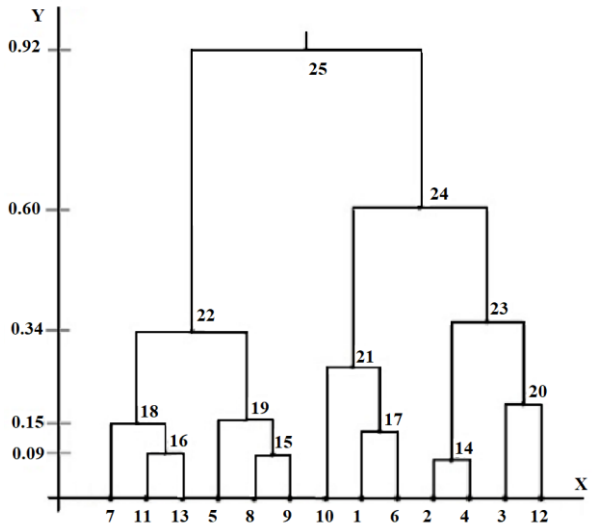

Fig. 3. Dendrogram of clustering a group of 13 objects into subgroups.

The visual analysis of dendrogram shows that the objects are grouped into four clusters. Dendrogram visualization is easy to read because the hierarchy of nodes and associations has a natural and intuitive look. The numbering of nodes goes in the direction of growth according to their metric given by the ordinate. In the direction of the abscissa, the nodes are oriented in accordance with the unification strategy.

This dendrogram clearly identifies four clusters to which the following objects belong to:

- cluster 1 of objects 7, 11, 13;
- cluster 2 of objects 2, 3, 4, 12;
- cluster 3 of objects 5, 8, 9;
- cluster 4 of objects 1, 6, 10.

In addition, the distance between node 22 and node 18, 19 and between node 24 and node 21, 23 is bigger than distance between objects. It can be argued that these clusters are fairly well "isolated" from each other.

The constructed dendrogram for 20 objects from Table 2 is presented in Fig. 4. The rules for constructing the dendrogram above are also taken into account. Visual analysis of the dendrogram indicates that it can also distinguish four clusters with the following objects:

- cluster 1 - objects 20, 10, 12, 19, 1, 2;
- cluster 2 - objects 11, 8, 9, 6, 4, 3, 5;
- cluster 3 - objects 14, 17, 18;
- cluster 4 - 15, 16, 7, 13.

Further, visual analysis indicates that distance between closest objects in the clusters is at least twice bigger than distance between these objects and their respective groups. In addition, distance between 23 and 33 nodes is almost the same as that of the distance between the nearest objects. This can be explained by the fact that nodes in such a cluster are placed more tightly than objects in the cluster itself.

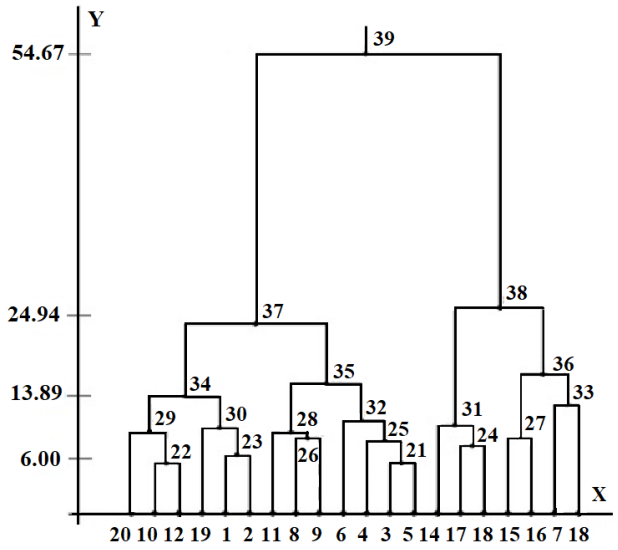

Fig. 4. Cluster dendrogram of a group of 20 objects.

In addition, it can be noted that the distance between node 37 and nodes 34 and 35 and between node 38 and between nodes 31 and 36 is much smaller than between node 39 and nodes 37 and 38. This indicates that clusters 1 and 2 are quite close to each other. Accordingly, clusters 3 and 4 are quite close too.

Thus, visual analysis of the above dendrograms allows us to underline some qualitative conclusions about the identified clusters. However, the information about density can be evaluated. It means, a cluster is a set of data objects spread in the data space over a contiguous region of high density of objects. Density can be explained as group of objects with minimum distance between each other. The distance in Fig. 4 is presented on Y-axis. Therefore, the closes objects are situated in cluster consists of objects 3, 5, 4, 6.

3.4. Disclosure of the dendrogram by hypothesis

It is suggested that a more detailed interpretation of the dendrogram is made by implementing the hypothesis given above. Testing this hypothesis means a procedure of visualizing and interpreting the result of a cluster analysis. The disclosed dendrogram allows for full usage of association metrics.

Since this metric is derived from the calculation of values of the proximity matrix in accordance with the chosen object pooling strategy, the use of the disclosed dendrogram is quite legitimate. In addition, the procedure for opening the dendrogram is specific and unambiguous.

The disclosure procedure is implemented by the following algorithm.

Algorithm 1. The disclosure of dendrogram

1. Objects are denoted by bold numbers and nodes by ordinary numbers.
2. Select in the center of a cluster field (rectangular area) the point at which we move the node that joins two largest clusters — that is, the root of the tree.
3. Draw a horizontal line through this point and measure the segments corresponding to the distance from this point to the left cluster node to the left side and from the right cluster node to the right in the selected scale.
4. Draw vertical lines through the ends of segments. Limit these lines (verticals) to distances to two corresponding clusters. Obviously, in the case of agglomerative hierarchical cluster analysis, these distances will decrease every time.

Alternating horizontal and vertical lines with specified lengths of distances between objects, between an object and a group, and between groups of groups, we obtain a detailed diagram of the dendrogram.

The images of clustering objects localized to their nodes in the selected scale will be obtained after the deployment of dendrogram on the plane. Here the ovals correspond to imaginary areas of selected clusters.

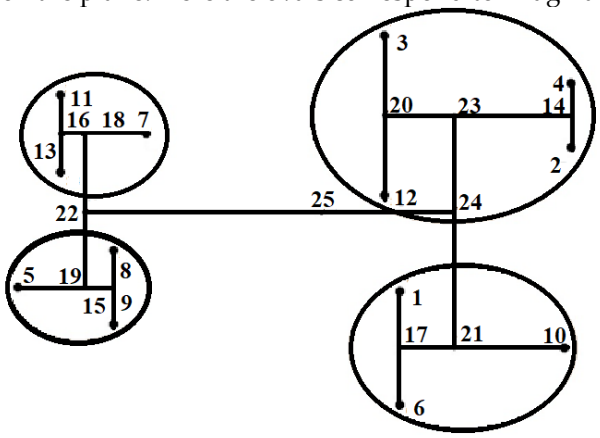

Fig. 5. The opening of the dendrogram from Fig. 3.

As a result of such representation in Fig. 4 it is possible to clearly identify the four well separate clusters. It can also be assumed that distances between objects in clusters are smaller than distances between clusters.

The disclosure of dendrogram shown in Fig. 4 is carried out according to the same algorithm. As a result, the open dendrogram has the form shown in Fig. 6.

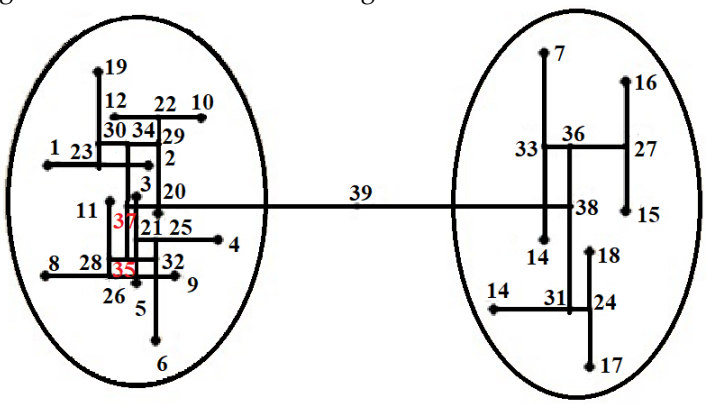

Fig. 6. Crohn's dendrogram of 20 objects.

Visual analysis of these dendrograms shows that ordinary dendrograms differ little from each other, but after their disclosure the difference between them is very significant. If, in the first case, objects can be clearly distinguished by four clusters, then in the second case, we can assume that there are only two clusters. Although the data structure in two examples is different, the dendrograms in both examples are the result of a well-defined mathematical procedure, namely the strategy of combining objects of attention.

Another important feature of the disclosed dendrograms is that they reflect the distribution of objects in the plane. Since such a plane shows result of the analysis, it can be called a cluster plane or a cluster plane. It is analogous to the Cartesian coordinate system, but with scales that correspond to the metric of associations.

The dimension of this plane on both axes has the same scales defined in the table of distances of associations. Since corresponding coordinates represent the objects in this plane, it is quite simple to calculate the perimeter and area.

Note that the ordinary dendrogram also lies in a rectangular plane, but its ordinate has a metric of union and abscissa lies in a nominal scale. Therefore, the peculiarity of open dendrograms is that their coordinates represent objects in clusters in a rectangular system. In addition, the numerical values of the abscissa and ordinate scales correspond to the metric used for the distance between objects.

3.5. Quantitative characteristics of conventional dendrograms.

Cluster analysis is not limited to identifying clusters, their number, and the affiliation of objects to a particular cluster. The main result is that different quantitative characteristics of clusters themselves and their objects can be calculated simultaneously.

We introduce the concept of the cluster center.

Definition 1. We consider a cluster center or central cluster node to be a node that covers only all objects in that cluster. By hierarchy, it is the highest node that integrates all cluster objects.

Based on the data in the aggregation table, the following characteristics can be calculated for these clusters. Therefore, for the dendrograms in Fig. 3 we have the following indicators.

(a) The average distance from cluster objects to the center node.

For the first cluster, this distance is defined as follows. The center of cluster 1 is node 18. The average distance is equal to the average of the following paths (segments):

- from object 7 to node 18;
- from object 11 to node 16 and from node 16 to node 18;
- from object 13 to node 16 and from node 16 to node 18.

Thus, the average distance for cluster 1 is equal to:

$$l_1 = \left[2 \cdot (0.094 + 0.058) + 0.152 \right] / 3 = 0.152.$$

Similarly, for centers: cluster 2 center node 19, cluster 3 node 23, cluster 4 node 21.

$$l_2 = \left[2 \cdot (0.09 + 0.072) + 0.162 \right] / 3 = 0.162,$$

$$l_3 = \left[2 \cdot (0.195 + 0.168) + 2 \cdot (0.077 + 0.286) \right] / 4 = 0.363,$$

$$l_4 = \left[2 \cdot (0.139 + 0.131) + 0.270 \right] / 3 = 0.270.$$

The distance from the root to the central node is equal to the length of segments on the path from the root to the center of the cluster, for example:

- for cluster 1 it is equal c1=0.579+0.193=0.772;
- for cluster 2 it is equal c2=0.579+0.183=0.762;
- for cluster 3 it is equal c3=0.324+0.237=0.561;
- for cluster 4 it is equal c4=0.324+0.330=0.654.

The distance between the cluster centers. The values for the four clusters will be 6 (all possible combinations):

- between cluster 1 and cluster 2:
  S_{12}=0.183(node 18)+0.193(node 19)=0.376;
- between cluster 1 and cluster 3:
  S_{13}=0.183(node 18)+0.679 (node 24)+0.324(node 21)+0.237 (centroid of node 18 and node 21)=1.423;
- between cluster 1 and cluster 4:
  S_{14}=0.183 (node 18)+0.679 (node 24)+0.344 (node 23)+0.330 (centroid of node 18 and node 24)=1.536;
- between cluster 2 and cluster 3:

S_{23}=0.193(node 19) +0.679(node 24)+0.324(node 21)+0.337 (centroid of node 19 and node 24)=1.533;

- between cluster 2 and cluster 4:
  S_{24}=0.193(node 19)+0.679(node 24)+0.344(node 23)+0.330(centroid)=1.546;
- between cluster 3 and cluster 4:
  S_{12}=0.324(node 21)+ 0.344(node 23)=0.668.

There are four clusters in the dendrogram in Fig. 4. A horizontal line that runs below node 37 but above node 36 distinguishes them. In this case, the center of the first cluster is node 34, the second is 35, the third is 31, and the fourth is 36. The characteristics of these clusters are as follows.

(b) The average distance from cluster objects to the center node.

For the first cluster, this distance is defined as follows. The center of cluster 1 is node 34. The average distance is equal to the average of the following paths (segments):

- from object 20 to node 29 and from node 29 to node 34 and from node 34 to node 37;
- from objects 10 or 12 to node 22 and from 22 to 29 and from 29 to 34 and from 34 to 37;
- from object 19 to node 30 and from node 30 to 34 and from node 34 to node 37.

Thus, the average distance from the objects to their central node 34 is equal. Similarly, for centers:

- cluster 1 center node 34 - l_1=13.89;
- cluster 2 center node 35 - l_2=15.43;
- cluster 3 center node 31 – l_3=10.5;
- cluster 4 center node 36 – l_4=16.59.

The distance from the root to the central node — it is equal to the length of the segments on the path from the root to the center of the cluster, for example:

- for cluster 1 it is equal c_1=40.78;
- for cluster 2 it is equal c_2=39.24;
- for cluster 3 it is equal c_3=44.17;
- for cluster 4 it is equal c_4=38.08.

The distance between the cluster centers — the values for the four clusters will be 6:

- between cluster 1 and cluster 2 d_{12}=15.7;
- between cluster 1 and cluster 3 d_{13}=84.95;
- between cluster 1 and cluster 4 d_{14}=78.86;
- between cluster 2 and cluster 3 d_{23}=83.41;
- between cluster 2 and cluster 4 d_{24}=77.32;
- between cluster 3 and cluster 4 d_{12}=22.79.

Obviously, the use of these indicators is determined by the objectives set. However, even in such representation, indicators can be used to compare the results of cluster analysis by different methods.

4. Results

4.1. Own method.

The real and convex shapes of cluster regions for both exposed dendrograms are shown in Fig. 7 and Fig. 8.

The visual analysis of these figures indicates that the shape of the real cluster areas can be very different. The area of the cluster is regarded as a polygon. The boundary of a cluster region is a broken line that consistently connects all the objects of the cluster without intersecting. This is a real form of cluster area. For the convex shape of the cluster region, only the localization (coordinates) of those objects that provide the convexity of the cluster region are taken into account. In other words, the convex shape of the

cluster is also a broken line, but it must satisfy the convexity condition of the polygon. Obviously, the value of the area in this case will be overstated. However, the convex region is strictly defined and is not arbitrary.

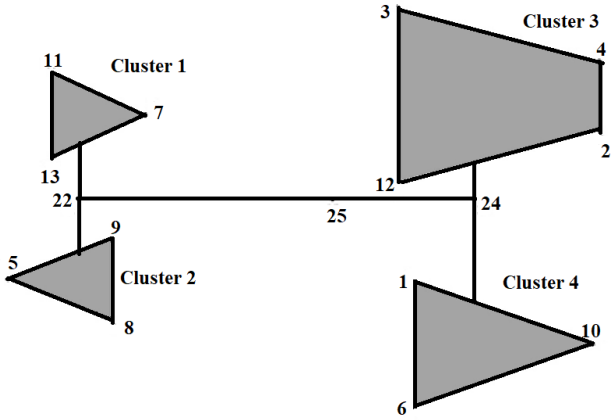

Fig. 7. Form of clusters for distribution of 13 objects

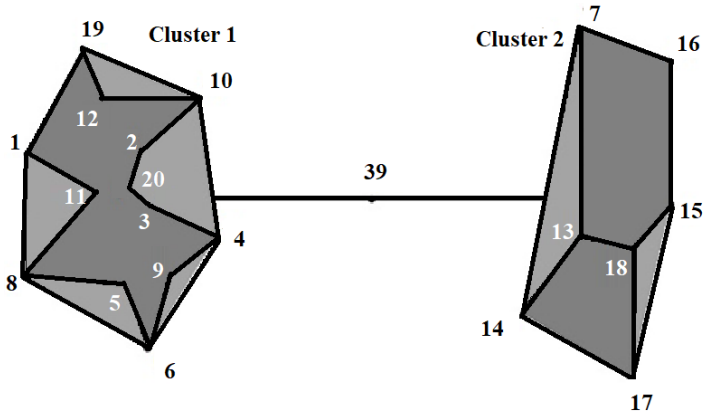

Fig. 8. Form of clusters for distribution of 20 objects

Presentation of the open dendrogram as localization of objects on a plane gives the following indicators:

1. The values of the coordinates of the objects in the cluster plane.
2. The value of the linear distance between the objects and between the centers of the clusters.
3. The value of the real cluster area.
4. The value of the area of the minimum convex area of the cluster coverage.
5. The value of the ratio of the perimeter of the cluster area to the square root of its area.
6. The value of the average area attributable to one object of the cluster.

In Fig. 7 dendrogram clusters for a sample of 13 objects are well separated on the plane. The shape of the areas of these clusters is simple and corresponds to simple geometric shapes. In Fig. 8 real and convex cluster shapes for 20 objects are presented. The lighter color highlights the minimally convex and the darker the real shape of the cluster regions.

The disclosed dendrogram lies in a plane with both axes having the same scales; as defined in the table of distances of associations. This, in turn, gives grounds for determining the perimeter and area of cluster areas. Since the objects in this plane are represented by the corresponding coordinates, it is quite simple to calculate the perimeter and area. Table 3 shows the results of calculating the perimeter and cluster area for

both disclosed dendrograms. In addition, perimeters for different shapes of cluster regions are also calculated.

The perimeter and area indicate the close relationship between the objects in the cluster. Obviously, the smaller the object area, the denser the objects localized in the cluster. The ratio of the cluster area to the number of objects can be used to rank clusters in a single sample. In this case, the smaller the value, the higher the cluster rank.

Table 3. Characteristics of clusters of two samples

Sampling of 13 objects				
Clusters	**1**	**2**	**3**	**4**
Perimeter	324.078	340.711	742.523	563.027
Square	4935	5265	31008	13869
Parameter	4.613	4.695	4.217	4.781
Density	1645	1755	7752	4623
Sampling of 20 objects				
Clusters	**1**		**2**	
The shape of the area	Real	Convex	Real	Convex
Perimeter	663.745	583.058	642.359	624.905
Square	12784	22822	14924	19640
Parameter	5.870	3.859	5.258	4.459
Density	983.4	1755.5	2132	2805.7

The perimeter-to-root ratio of the area is also an effective and generalized indicator. The fact is that no matter what shape the cluster is, it is a polygon. Therefore, the ratio of the perimeter to the root of the square does not depend on either the shape or the size. In addition, since there are known coordinates of vertices of the polygon and metric scales, the perimeter and area can be calculated by known formulas.

4.2. Dendrogdam visuzlization using standart approaches.

Dendrogram visualization in 3d plane can be done using standard approach. For example, we can create the ensemble from hierarchical clustering and Principal Component Analysis (PCA). FactoMineR is a R package which is developed in AgrocampusOuest, and is dedicated to factorial analysis. The aim is to create a complementary tool for this package which is dedicated to clustering. This is created after the factorial analysis [16].

Fig. 9 represents the result of factorial analysis for second dataset. Let us evaluate the quality of this method. The dataset is well visualized with strict division on clusters. The time complexity can be estimated for PCA and hierarchical clustering separately. Covariance matrix computation is $O(p^2n)$, where p is number of features and n is number of instances; its eigen-value decomposition is $O(p^3)$. So, the complexity of PCA is $O(p^2n+p^3)$. The standard algorithm for hierarchical agglomerative clustering has a time complexity of $O(n^3)$. However, for single-linkage with a heap the runtime of the general case can be reduced

to $O(n^2 \log n)$ [17]. So, time complexity of factorial analysis is $O(p^2 n + p^3 + n^2 \log n)$.

The overall time complexity of the disclosure algorithm (Algorithm 1) is $O(n^2 \log n + pn)$. Time complexity $O(n^2 \log n)$ is used for dendrogram building and complexity $O(pn)$ is used for each feature drawing and disclosure.

Therefore, time complexity for proposed algorithm is significantly less than for factorial analysis.

The next important thing, factorial analysis presents the dataset in 3d without perimeter and square evaluation.

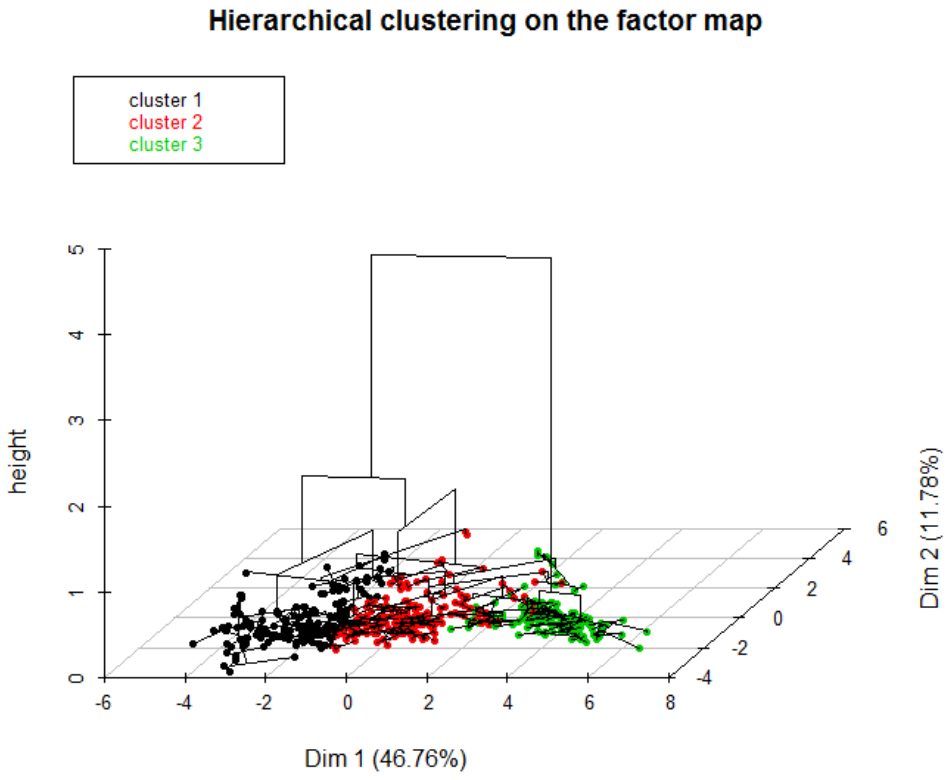

Fig. 9. Factorial analysis for second dataset.

Discussion

The result of the research and verification of the proposed hypothesis is the disclosure of the dendrogram algorithm as the extension of classical methods of cluster analysis. This extension is made by studying and disclosing the resulting image of the dendrogram. The dendrogram visualization thus obtained differs significantly from the classical results. The opening of the dendrogram according to the developed algorithm allows us 3D visualization of the analysis results, as well as calculating the area and perimeter of the obtained clusters. Therefore, using analytical geometry methods, it is quite easy to isolate and calculate the parameters of minimum cluster coverage surfaces and the immediate distances between any objects of one or different clusters, as well as between the objects of a given cluster. This, in turn, is a significant complement to cluster analysis.

Generally, the dendrogram is constructed based on calculated distances between objects, object and group, and between two groups in accordance with a specific strategy of unification. The opening of the dendrogram for the second dataset clearly indicates that the objects of the first and second clusters are quite close to each other. It can also be assumed that the clusters are built without intersection.

Comparing the dendrograms in both cases, it is easy to see the difference in merging distances in the first and second steps. For the first case, the object-object and object-group combining distances are significantly smaller than the same distances in the second. This means that in the first case, the objects are more closely related. In the second case, objects may also appear in the adjacent cluster.

This paper does not compare our results with computational intelligence algorithms like monarch butterfly optimization (MBO), earthworm optimization algorithm (EWA), elephant herding optimization (EHO), moth search (MS) algorithm. These algorithms are used mostly with iterative clustering algorithms such as k-means. Optimal positions of the centroids are searched by optimization algorithm. However, the purpose of the paper is modification of noniterative clustering methods such as agglomerative hierarchical clustering.

Conclusion

The main contribution of the paper is to develop a dendrograms-based method for 3D visualization of hierarchical clustering for multidimensional data. This method is built on hierarchical clustering algorithm. The overall time complexity of the disclosure algorithm is $O(n^2 \log n + pn)$, which is on pn bigger than for single-link agglomerative clustering. The dendrogram in proposed method is presented as a 3D tree, the object line is corresponded to the cross-sectional plane of the branches. The main advantages of proposed algorithm include: (a) areas of the selected clusters and their objects on a plane can be calculated, according to the coordinates defined by the open dendrogram; (b) the rules for visualization of the dendrogram are defined which allow to find the nature of clusters; (c) the quantitative indicators of localization of objects and evaluation of clusters are proposed.

Perimeter and area indicators indicate the closeness of the connection of objects in the cluster. That is, that the smaller the area per object, the more densely localized the objects in the cluster. The ratio of the cluster area to the number of objects in it can be used for ranking clusters. In this case, the smaller this value, the higher the rank cluster.

The proposed method is evaluated using different datasets. The proposed method significantly improves the quality of visualization and evaluation of cluster analysis results. It is also efficient as the time complexity is significantly less for factorial analysis.

The disclosed dendrogram retains proportions in distances between objects. On the basis of these characteristics, it is possible to determine the close relationship between the clusters themselves by correlating the values of their quantitative averaged values of the traits.

Thus, the opening of the dendrogram allows us to clearly identify the set of clusters, each of which has its own distribution of the range of features values. The quantitative characteristics of clusters on both dendrograms are quite simple. In addition, the mean values of the features of objects in a given cluster can be interpreted as generalized characteristics of this cluster, and the cluster itself can be represented as a single integral object.

References

1. Waibel, Christoph; EVINS, Ralph; CARMELIET, Jan. Clustering and Ranking Based Methods for Selecting Tuned Search Heuristic Parameters. In: 2019 IEEE Congress on Evolutionary Computation (CEC). IEEE, **2019**: 2931-2940.
2. Bongiorno, Christian; Miccichè, Salvatore; Mantegna, Rosario N. Nested partitions from hierarchical clustering statistical validation. arXiv preprint arXiv:1906.06908, **2019**

3. Kaminskyi, Roman, et al. Methods of statistical research for information managers. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). IEEE, **2018**. p. 127-131
4. T. Soni Madhulatha. An Overview on Clustering Methods. IOSR Journal of Engineering Apr. **2012**, 2(4): 719-725.
5. Dimitar L.Vandev, Yanka G.Tsvetanova. About Ordering Features of Single Linkage Clustering Algorithm. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.768&rep=rep1&type=pdf
6. Trivedi, Shrawan Kumar, et al. Handbook of research on advanced data mining techniques and applications for business intelligence. IGI Global, **2017**, 122 p.
7. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis. On Clustering Validation Techniques. Journal of Intelligent Information Systems, **2001**, 17:2/3, 107–145
8. Szczuciński, P. Analiza skupień w badaniu struktury funkcjonalnej gmin na przykładzie województwa lubuskiego. Turystyka i Rozwój Regionalny, **2017**, 105-114.
9. Řezanková, Hana. Cluster Analysis of Economic Data. University of Economics. Prague, Czech Republic STATISTIKA, **2014**: 94-100.
10. Lytvyn, Vasyl, et al. A Smart Home System Development. In: International Conference on Computer Science and Information Technology. Springer, Cham, **2019**. p. 804-830.
11. Govender, P., and V. Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980-2019)." Atmospheric Pollution Research, **2019**
12. Lin, X. A Road Network Traffic State Identification Method Based on Macroscopic Fundamental Diagram and Spectral Clustering and Support Vector Machine. Mathematical Problems in Engineering, **2019**, 312 p.
13. Cohen-Addad, Vincent, et al. Hierarchical clustering: Objective functions and algorithms. Journal of the ACM (JACM) 66.4, **2019**, 1-42.
14. Kassambara, Alboukadel. Practical guide to cluster analysis in R: Unsupervised machine learning. Vol. 1. STHDA, 2017. https://www.datanovia.com/en/product/practical-guide-to-cluster-analysis-in-r/
15. Dimitar L.Vandev, Yanka G.Tsvetanova. Ordered Dendrogram. Available from: https://store.fmi.uni-sofia.bg/fmi/statist/personal/vandev/papers/orden1.pdf
16. Beautiful dendrogram visualizations in R: 5+ must known methods - Unsupervised Machine Learning. Available from: http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning
17. Naseem, Rashid, et al. Euclidean space based hierarchical clusterers combinations: an application to software clustering. Cluster Computing 22.3, **2019**, 7287-7311
18. Huang, S., Cui, H., Wei, X., & Cai, Z. Clustering-Based Monarch Butterfly Optimization for Constrained Optimization. International Journal of Computational Intelligence Systems, 13(1), 2020, 1369-1392.
19. Tuba, E., Dolicanin-Djekic, D., Jovanovic, R., Simian, D., & Tuba, M. Combined with K-means for data clustering. In Information and Communication Technology for Intelligent Systems 2019, 665-673.
20. Shukla, R. M., & Sengupta, S. Scalable and Robust Outlier Detector using Hierarchical Clustering and Long Short-Term Memory (LSTM) Neural Network for the Internet of Things. Internet of Things, 2020, 100167.
21. Byung Wan Jo, Rana Muhammad Asad Khan, Yun Sung Lee, Jun Ho Jo, Nadia Saleem. A Fiber Bragg Grating-Based Condition Monitoring and Early Damage Detection System for the Structural Safety of Underground Coal Mines Using the Internet of Things. Journal of Sensors, vol. 2018, Article ID 9301873, 16 pages, 2018. https://doi.org/10.1155/2018/9301873
22. MavLab Sensor data: http://casas.wsu.edu/datasets/mavlab.zip