Ministry of Education and Science of the Russian Federation

Federal State Autonomous Educational Institution of Higher Professional Education

Saint Petersburg National Research University

of Information Technologies, Mechanics and Optics

Faculty of Institute of Urban Studies and Design

REPORT

about Course work

The task title: «Clusterization of social network users by gender»

Student

Khomyakov VV C4111c

(Surname, initials) Group

Supervisor: Bochenina K. O., Ph.D.

Saint-Petersburg, 2019

CONTENT

TASK STATEMENT 3

1. OVERVIEW 4

2. ALGORITHMS 6

2.1 k-Means algorithm 6

2.2 Kohonen network 7

2.3 EM algorithm 9

2.4 Hierarchical clustering 10

3. EXPERIMENTAL RESEARCH 11

3.1 Primary analysis 11

3.2 K-means 12

3.3 Kohonen Self-Organising Maps (SOM) 14

3.4 Gaussian Mixture 15

3.5 Hierarchical clustering 16

CONCLUSION 17

REFERENCE 18

3Ошибка! Источник ссылки не найден.

TASK STATEMENT

The task of the hackathon of MF TINT (date of the hackathon is 02.02.2019 - 13.02.2019)

According to the presented data set prepared on the basis of open pages of social networks, it is necessary to solve the problem of recognizing the sex of the page owner. The data array with a size of 9380 records contains neither message texts nor user identifiers — only an impersonal set of 14 signs characterizing the behavior of each individual user is available.

The gender label is recorded in numerical form: 0 for female and 1 for male. The assessment of the quality of the task will be made on the basis of the share of correctly recognized tags.

The names of the fields in the provided file are interpreted as follows:

ID - the local user ID in the data set;

Followers_count - the number of followers;

Friends_count - the number of friends the user has;

Wall_comments - are there any comments on the user's wall? ;

Comment_count - number of comments left by user;

Post_count - the number of posts published by the user;

Reposts_count - the number of reposts published by the user;

Verified - whether the account is verified;

Videos_count - the number of published user videos;

Photos_count- number of photos published by user;

Gifts_count - the number of gifts received by the user;

Sex - hidden column labeled gender (1 - male, 0 - female) ;

1. OVERVIEW

Classification is the simplest and most common task. As a result of solving the classification problem, signs are found that characterize groups of objects of the studied data set - classes; on these grounds, a new object can be attributed to a particular class.

The task of clustering is similar to the task of classification, it is a logical continuation, but its difference is that the classes of the studied data set are not predefined. Thus, clustering is intended for splitting a set of objects into homogeneous groups (clusters or classes). If the sampled data are presented as points in the attribute space, then the clustering problem is reduced to the definition of “point condensations”.Целью кластеризации является поиск существующих структур.

Clustering is a descriptive procedure, it does not draw any statistical conclusions, but it makes it possible to conduct an exploratory analysis and study the "data structure". Clustering is a descriptive procedure, it doesn’t draw any data structure.

The characteristics of the cluster can be called two signs:

- internal homogeneity;
- external isolation.

Figure 1 shows the general clustering scheme.

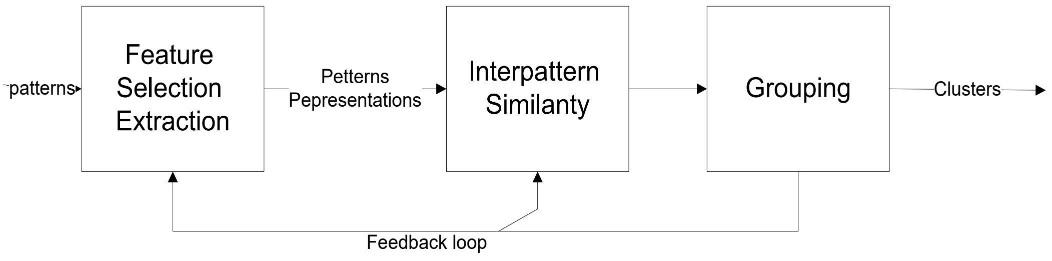

Figure 1 - General clustering scheme

Data clustering includes the following steps:

a) Selection of characteristics.

First you need to choose the properties that characterize our objects, they can be quantitative characteristics (coordinates, intervals ...), qualitative characteristics (color, status, military rank ...), etc. Then you should try to reduce the dimension of the space of characteristic vectors, that is, to highlight the most important properties of objects. Reducing the dimensions speeds up the clustering process and in some cases allows you to visually evaluate the results. The selected characteristics should be normalized. Further, all objects are represented as characteristic vectors. We will fully identify the object with its characteristic vector.

b) Metric Definition.

The next stage of clustering is the choice of a metric by which we will determine the proximity of objects. Metric is selected depending on:

– the space in which the objects are located;

– implicit cluster characteristics.

c) Presentation of the results.

The results of clustering should be presented in a convenient form for processing in order to assess the quality of clustering. Usually one of the following methods is used:

– representation of clusters by centroids;

– representation of clusters by a set of characteristic points;

– representation of clusters by their limitations.

Different clustering methods can create different clusters, and this is normal. However, the creation of similar clusters by different methods indicates the correctness of clustering.

Clustering methods. There are the following classification of approaches:

Probabilistic approach. It is assumed that each object will be, belongs to one of the k classes:

- K-medium;

- K-median;
- EM-algorithm;
- Algorithms of the FOREL family;
- Discriminant analysis.

Approaches based on the use of artificial intelligence: a conditional group, since there are so many methods, and they are quite different:

- The method of fuzzy clustering C-means (C-means);
- Kohonen's neural network;
- Genetic algorithm.

The logical approach. Building a dendrogram using a decision tree.

Graph-Theoretic approach:

- Graph Clustering Algorithms.

Other methods:

- Statistical clustering algorithms;
- Ensemble of clusterizers;
- Algorithms of the KRAB;
- Algorithm based on the screening method.

DBSCAN and etc.

2. ALGORITHMS

2.1 k-Means algorithm

Based on the task condition [1],[2], the system is asked to form exactly two clusters so that they are as different as possible. This is exactly the type of tasks that the k-means method algorithm solves. In the general case, the method builds exactly K different clusters located at as large distances from each other as possible.

Description of the algorithm.

The initial distribution of objects in clusters.

The number k is chosen, and at the first step these points are considered as "centers" of clusters. Each cluster corresponds to one center.

The selection of initial centroids can be done as follows:

- the choice of k-observations to maximize the initial distance;

- random selection of k-observations;

- selection of the first k-observations.

As a result, each object is assigned to a specific cluster.

Iterative process.

Cluster centers are calculated, which are then considered to be the coordinate coordinate clusters. Objects are redistributed again.

The process of calculating the centers and the redistribution of objects continues until one of the following conditions is fulfilled:

- cluster centers have stabilized, i.e., all observations belong to the cluster to which they belonged prior to the current iteration;

- the number of iterations is equal to the maximum number of iterations.

Advantages of the k-means algorithm:

- ease of use;

- speed of use;

- clarity and transparency of the algorithm.

Disadvantages of the k-means algorithm:

- The algorithm is too sensitive to outliers that can distort the average. A possible solution to this problem is to use a modification of the algorithm — the k-median algorithm;

- The algorithm may work slowly on large databases. A possible solution to this problem is to use data sampling.

2.2 Kohonen network

Kohonen's network is one of the varieties of neural networks that use unsupervised learning. With such training, the training set consists only of the values of the input variables; in the learning process, there is no comparison of the outputs of the neurons with the reference values.

The term “Kohonen network” was introduced in 1982 by the Finnish scientist Toivo Kohonen. Kohonen networks are a variety of self-organizing feature maps that, in turn, are a special type of neural network.

The main goal of Kohonen networks is the transformation of complex multidimensional data into a simpler structure of small dimension. Thus, they are well suited for cluster analysis, when you want to detect hidden patterns in large data arrays..

By the ways of setting the input weights of adders and by the tasks being solved, there are many types of Kohonen networks. The most famous of them are:

- Networks of vector quantization of signals, closely connected with the simplest basic cluster analysis algorithm (the method of dynamic nuclei or K-means);
- Kohonen self-organizing maps (Self-Organising Maps, SOM) [3]
- Learning Vector Quantization.

Kohonen network consists of nodes that are combined into clusters. The closest nodes correspond to similar objects, and distant from each other - unlike.

The basis of building the Kohonen network is competitive learning, when the output nodes (neurons) compete with each other for the right to become a “winner”

Network training. Consider a set of *m* field values of the *n*-th record of the original sample, which will serve as the input vector $X_n = (x_{n1}, x_{n2}, ..., x_{nm})^T$, and the current weight vector of the *j*-th output neuron $W_j = (w_{1j}, w_{2j}, ..., w_{mj})^T$.

Kohonen's algorithm includes the following steps:

1. Initialization. For neurons of the network, the initial weights are established, and the initial learning rate ŋ and the learning radius R are also set.

2. Excitement. An input vector X_n is fed to the input layer containing the values of the input fields of the training sample record.

3. Competition. For each output neuron, the distance $D(W_j, X_n)$ between the weight vectors of all the neurons of the output layer and the input action vector is

calculated. If the Euclidean distance is chosen as a measure of proximity of two vectors, then we get

$$D(W_j, X_n) = \sqrt{\sum_i \left(w_{ij} - x_{ni}\right)^2} \tag{1}$$

In other words, the distance between the weights of all neurons of the output layer and the vector of input is calculated. That neuron j, for which the distance will be the smallest, and will be the winner.

4. Combining. All neurons located within the learning radius relative to the winning neuron are determined.

5. Adjustment. The adjustment of the neuron weights within the learning radius is made in accordance with the formula of a linear combination of input vectors and current weights of the weights:

$$w_{i,j,new} = w_{i,j,current} + \eta\left(x_{ni} - w_{i,j,current}\right) \tag{2}$$

At the same time, the weights of the neurons closest to the winning neuron are adjusted in the direction of its weight vector.

6. Correction. The radius and parameter of the learning rate change in accordance with a given law [4].

2.3 EM algorithm

The EM (Expectation-Maximization) algorithm is based on the calculation of distances, i.e. detecting areas that are more “populated” than others. In the process of the algorithm, an iterative improvement of the solution occurs, and the stop is made at the moment when the required level of accuracy of the model is reached. [5]

The EM algorithm is based on the assumption that the data set under investigation can be modeled using a linear combination of multidimensional normal distributions. It is assumed that the data in each cluster obey a certain distribution law, namely, the normal distribution.

The EM algorithm is an iterative algorithm, each iteration of which consists of two steps: the expectation step (E-step) and the maximization step (M-step).

At the E-step, the expected value of the likelihood function is calculated, and the hidden variables are treated as observables. At the M-step, the maximum likelihood estimate is calculated, so the expected likelihood, which is calculated at the E-step, increases. This value is then used for the E-step at the next iteration. The algorithm runs until convergence. We describe these steps from a mathematical point. To do this, consider the function:

$$F(q,\theta) = E_q[\log L(\theta;x;Z)] + H(q) = -D_{KL}(q|\left|p_{z|x}(\cdot|x;\theta)\right) + \log L(\theta;x) \quad (3)$$

where, q – probability distribution of unobservable variables Z; $p_{Z|X}(\cdot \mid x;\theta)$ – conditional distribution of unobservable variables with fixed observables x and parameters θ; H – entropy; D_{KL} — distance.

Then the steps of the EM algorithm can be represented as:

E-step: Select q to maximize F:

$$q^{(t)} = \arg\max_q F(q,\theta^{(t)}) \quad (4)$$

M-step: Choose θ, to maximize F:

$$\theta^{(t+1)} = \arg\max_\theta F(q^{(t)},\theta) \quad (5)$$

2.4 Hierarchical clustering

The essence of hierarchical clustering consists in successively merging smaller clusters into larger ones (agglomerative methods) or dividing large clusters into smaller ones (divisional methods) [5],[6].

Hierarchical agglomerative methods (Agglomerative Nesting, AGNES) are characterized by a consistent union of the original elements and a corresponding decrease in the number of clusters. At the beginning of the algorithm, all objects are separate clusters. In the first step, the two most similar objects are combined into a cluster. In subsequent steps, the merge continues until all the objects form one cluster.

Hierarchical divisional (divisible) methods (DIvisive ANAlysis, DIANA) are the logical opposite of agglomerative methods. At the beginning of the algorithm, all objects belong to the same cluster, which is divided into smaller clusters in subsequent steps, resulting in a sequence of splitting groups.

The advantage of hierarchical clustering methods is their visibility. However, hierarchical cluster analysis methods are used with small data sets.

Hierarchical algorithms are associated with the construction of dendrograms (from the Greek dendron - "tree"), which are the result of hierarchical cluster analysis. Dendrogram describes the proximity of individual points and clusters to each other, graphically represents the sequence of combining (dividing) clusters.

A dendrogram (dendrogram) is a tree diagram containing levels, each of which corresponds to one of the steps of the process of sequential integration of clusters. A dendrogram is also called a tree, a cluster merging tree, a tree of a hierarchical structure. A dendrogram is a nested grouping of objects that changes at different levels of the hierarchy.

3. EXPERIMENTAL RESEARCH

3.1 Primary analysis

We will conduct the primary data analysis:

1) Also on the constructed histograms it can be seen that in several columns there are zero data exceeding half of all values in the column:

   zero values in id = 0.0 %

   zero values in followers_count = 2.217956920452122 %

   zero values in comment_count = 0.0 %

   zero values in post_count = 0.0 %

   zero values in like_count = 0.0 %

   zero values in friends_count = 0.0 %

```
zero values in  group_count  =  32.341650671785025 %
zero values in  posts_count  =  1.599488163787588 %
zero values in  reposts_count  =  1.034335679249307 %
zero values in  videos_count  =  3.625506504585199 %
zero values in  audios_count  =  53.28428236297717 %
zero values in  photos_count  =  1.183621241202815 %
zero values in  gifts_count  =  59.68223501812753 %
```

Then, it makes sense to remove columns from the general table that have more than 30% of zeros

2) Remove non-role columns – ‘Id’ and ‘Verified’
3) Normalize the data. To do this, we use the normalization from the package python - sklearn.

3.2 K-means

We use a ready-made ready-made algorithm in the sklearn package. At the same time, we know that the number of clusters is 2.

```
kmeans = sklearn.cluster.KMeans(n_clusters=2)
kmeans.fit(data_array)
```

Calculate the number of found values for each cluster and their percentage. This is shown in the figure 2.

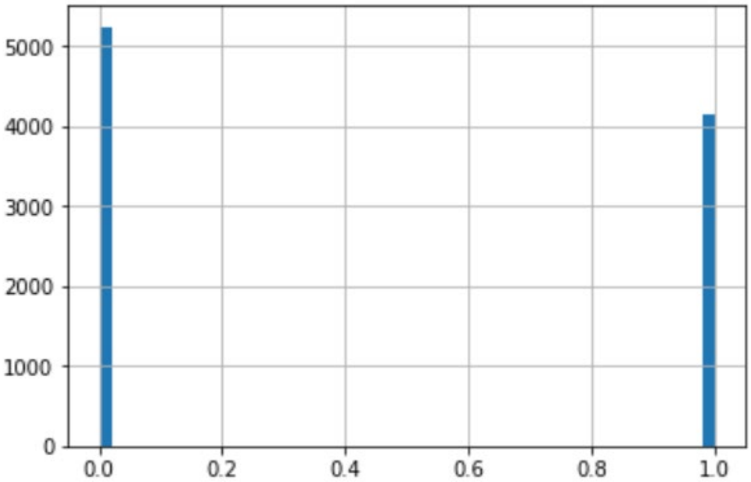

Figure 2. Histogram of the distribution of values across clusters. For cluster “0”, 55.9% of all values were detected, for cluster “1” - 44.1%.

We will evaluate clustering on the basis of two values: Silhouette Coefficient [7] and Davies-Bouldin score [8].

The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:
a: The mean distance between a sample and all other points in the same class.
b: The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b-a}{\max(a,b)} \tag{6}$$

The Davies-Bouldin index can be used to evaluate the model, where a lower Davies-Bouldin index relates to a model with better separation between the clusters. The index is defined as the average similarity between each cluster C_i for $i=1,...,k$ and its most similar one C_j. In the context of this index, similarity is defined as a measure R_{ij} that trades off:

s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.
d_{ij}, the distance between cluster centroids i and j.

A simple choice to construct R_{ij} so that it is nonnegative and symmetric is:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{7}$$

Then the Davies-Bouldin index is defined as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{8}$$

Zero is the lowest possible score. Values closer to zero indicate a better partition.

Using the sklearn package for clustering with the Kmeans algorithm, we obtained the following values:

- The Silhouette Coefficient is 0.30888;
- The Davies-Bouldin score is 1.4801.

3.3 Kohonen Self-Organising Maps (SOM)

When clustering by this method, we used a ready-made computational package for networks: neupy.

By setting the epochs parameter to 200, we achieved the following results (we use the same data representations as in 3.1):

- For cluster “0”, 58.2% of all values were detected;
- For cluster “1” - 41.8%;
- The Silhouette Coefficient is 0.30995;
- The Davies-Bouldin score is 1.49143.

Figure 3 shows the clusters obtained by reducing the dimension of space by the method of principal components.

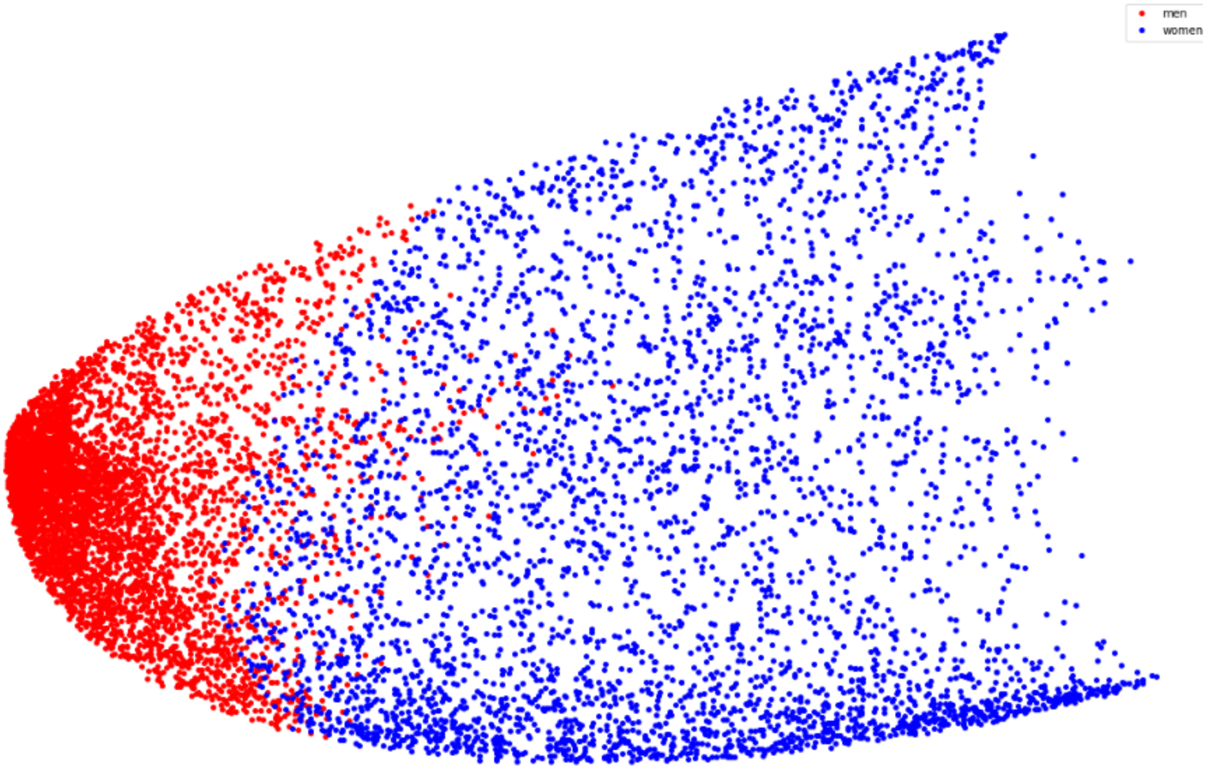

Figure 3. Cluster image obtained by decreasing the space dimension

3.4 Gaussian Mixture

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. The GaussianMixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models.

When using this method, the following results were obtained:

- For cluster “0”, 55.91% ;
- For cluster “1” - 44.09% ;
- The Silhouette Coefficient is 0.30888;
- The Davies-Bouldin score is 1.480096.

Reducing the dimension of space by the method of principal components, we obtain the visual distribution of clusters (Figure 4).

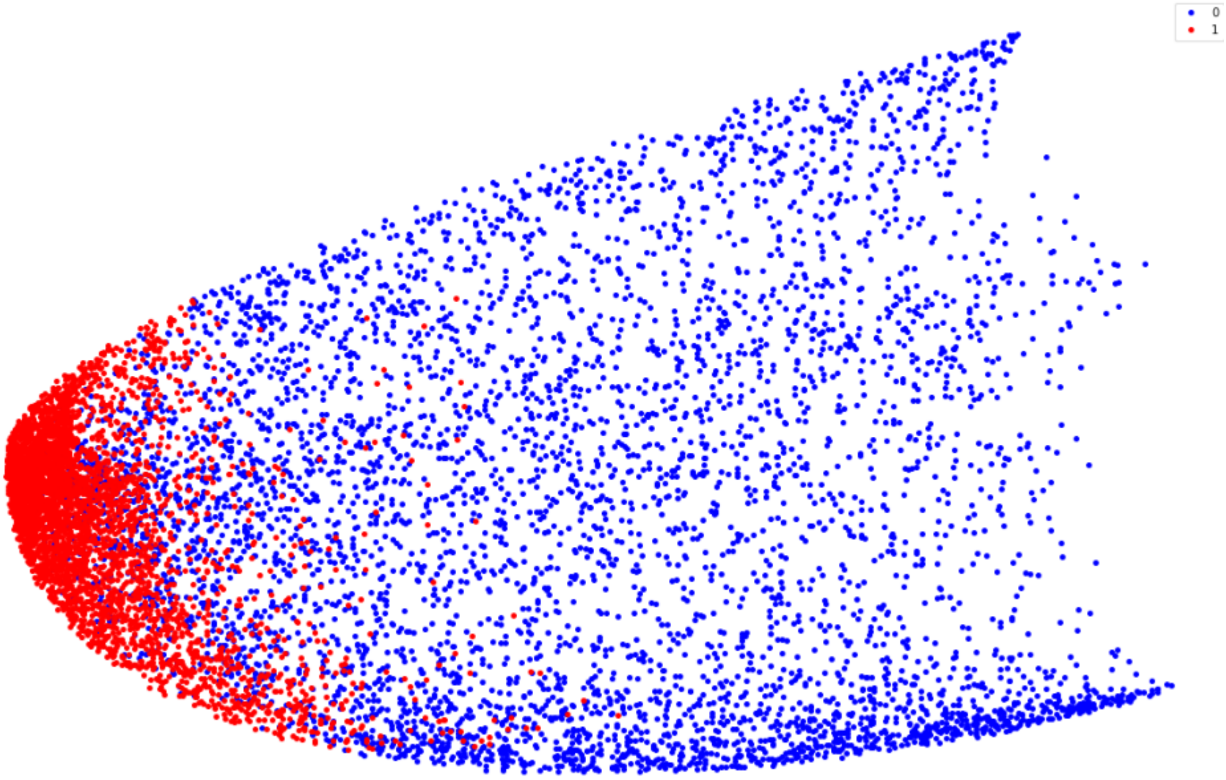

Figure 4. Cluster distribution obtained by EM clustering

3.5 Hierarchical clustering

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

Use Agglomerative Clustering. Hierarchical clustering is performed using a bottom-up approach: each observation begins in its own cluster, and the clusters are sequentially merged together. Communication criterion used: Ward, which minimizes the sum of squares of differences in all clusters.

Results:

- For cluster “0” - 52.804% ;
- For cluster “1” - 47.196% ;
- The Silhouette Coefficient is 0.27023;
- The Davies-Bouldin score is 1.59844.

CONCLUSION

Combine the research results (table 1):

Table 1 - the results of the study

	K-means	Kohonen Self-Organising Maps (SOM)	Gaussian Mixture	Hierarchical clustering
Доля кластера «0»	55.9%	58.2%	55.91%	52.804%
Доля кластера «0»	44.1%	41.8%	44.09%	47.196%
Silhouette Coefficient	0.30888	0.30995	0.30888	0.27023
Davies-Bouldin	1.4801	1.49143	1.480096	1.59844

In this work, the clustering of social network users by gender was carried out using several methods: k-means, SOM, EM and Hierarchical clustering.

Based on the results obtained, it can be judged that data clustering was successful, since all the clustering methods used gave a similar result. This is clearly seen in Figures 3 and 4, which shows the distribution of clusters.

In addition, clustering quality metrics were calculated - Silhouette Coefficient and Davies-Bouldin. Based on these readings, it can be assumed that clustering using Kohonen networks (SOM) gave the best result. When hierarchical clustering gave the worst result.

REFERENCE

[1] MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability – vol. 1. – pp. 281–297

[2] Steinhaus, H. Sur la division des corps matériels en parties // Bull. Acad. Polon. Sci. (in French). – T. 4. – Vol. 12. – P. 801–804. – 1957

[3] Kohonen, T. Self-Organizing Maps // New York: Springer-Verlag. – 2001

[4] Kohonen's neural network - (http://ru.wikipedia.org/wiki/Neuronal_network_Kohonen)

[5] Vorontsov, K.V. Algorithms for clustering and multidimensional scaling: studies. manual // K. V. Vorontsov - Moscow: Moscow State University, 2009. - 80 p.

[6] Bolshakova, E.I., Klyshinsky, E.S., Lande D.V., Automatic processing of texts in natural language and computational linguistics: studies. manual // Moscow: MIEM, – P. 272. – 2011.

[7] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65

[8] Davies, David L.; Bouldin, Donald W. A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI – Vol. 2. – P. 224–227. – 1979.