IRSTI 20.23.17 DOI: https://doi.org/10.26577/JMMCS2024-v123-i3-4

A.Kh. Nishanov[1*], M.Kh. Akbarova[1], A.T. Tursunov[2], F.F. Ollamberganov[1], D.E. Rashidova[3]

[1]Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, Uzbekistan, Tashkent

[2]Tashkent Pharmaceutical Institute, Uzbekistan, Tashkent

[3]Samarkand Institute of Economics and Service, Uzbekistan, Samarkand

[*]e-mail: nishanov_akram@mail.ru

CLUSTERING ALGORITHM BASED ON OBJECT SIMILARITY

The article examines the issue of drug clustering. Initially, k classes are arbitrarily formed and the resulting training sample is pre-processed, then the similarities between the objects of each class are evaluated based on the proximity function and the criterion for evaluating the contribution of objects to the formation of their own class. Usually, it is in percentage and is the degree of mutual similarity of objects of each class. In the next steps of the algorithm, first, one object is taken from the first class, and by adding it to all k classes, the contribution of this object to this class is measured. The object will be left in the class which has the most contribution. This process is repeated several times in a row for all objects of the class. The process is stopped when the location of objects does not change and the degree of similarity exceeds the required percentage. As a result, the required clusters are formed.

Key words: Clustering, proximity function, degree of similarity of objects, contribution of object to the class.

А.Х. Нишанов[1*], М.Х. Акбарова[1], А.Т. Турсунов[2], Ф.Ф. Олламберганов[1], Д.Э. Рашидова[3]

[1]Мұхаммед әл-Хорезми атындағы Ташкент ақпараттық технологиялар университеті, Өзбекстан, Ташкент қ.

[2]Ташкент фармацевтикалық институты, Өзбекстан, Ташкент қ.

[3]3Самарқанд экономика және сервис институты, Өзбекстан, Самарқанд қ.

[*]e-mail: e-mail: nishanov_akram@mail.ru

Объектің ұқсақтығына негізген кластерлік алгоритм

Мақалада дәрілік заттарды кластерлеу мәселесі қарастырылады. Бастапқыда k класы ерікті түрде қалыптасады және нәтижесінде алынған оқыту үлгісі алдын ала өңделеді, содан кейін жақындық функциясы мен объектілердің өз класының қалыптасуына қосқан үлесін бағалау критерийі негізінде әрбір сынып объектілерінің арасындағы ұқсастықтар бағаланады. Әдетте, ол пайызбен көрсетіледі және әр класс объектілерінің өзара ұқсастық дәрежесі болып табылады. Алгоритмнің келесі қадамдарында алдымен бірінші класстан бір объект алынады және оны барлық k классқа қосу арқылы осы объектінің осы классқа қосқан үлесі өлшенеді. Нысан ең көп үлес қосқан сыныпта қалады. Бұл процесс сыныптың барлық объектілері үшін қатарынан бірнеше рет қайталанады. Объектілердің орналасуы өзгермегенде және ұқсастық дәрежесі қажетті пайыздан асқанда процесс тоқтатылады. Нәтижесінде қажетті кластерлер қалыптасады.

Түйін сөздер: Кластер мәселесі, жақындық функциясы, объектілердің ұқсастық дәрежесі, объектінің сыныпқа қосқан үлесі.

А.Х. Нишанов[1*], М.Х. Акбарова[1], А.Т. Турсунов[2], Ф.Ф. Олламберганов[1], Д.Э.Рашидова[3]

[1]Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Узбекистан, г. Ташкент

[2]Ташкентский фармацевтический институт, Узбекистан, г. Ташкент
[3]Самаркандский институт экономики и сервиса, Узбекистан, г. Самарканд
*e-mail: nishanov_akram@mail.ru

Алгоритм кластеризации основанный на сходстве объектов

В статье рассматривается проблема кластеризации лекарственных средств. Первоначально произвольно формируются k классов и предварительно обрабатывается полученная обучающая выборка, затем оценивается сходство между объектами каждого класса на основе функции близости и критерия оценки вклада объектов в формирование собственного класса. Обычно он выражается в процентах и представляет собой степень взаимного сходства объектов каждого класса. На следующих шагах алгоритма сначала из первого класса берется один объект и путем добавления его ко всем k классам измеряется вклад этого объекта в этот класс. Объект останется в классе, внесшем наибольший вклад. Этот процесс повторяется несколько раз подряд для всех объектов класса. Процесс останавливается, когда расположение объектов не меняется и степень сходства превышает необходимый процент. В результате формируются нужные кластеры.

Ключевые слова: Задача кластеризации, функция близости, степень сходства объектов, вклад объекта в класс.

1 Introduction

Clustering algorithms based on the similarity of objects are used to find commonalities among data and divide them into clusters. These algorithms operate on similarity or distance criteria within the data. The basic philosophy of these algorithms is that objects within each cluster are similar, but objects between different clusters are different.

The authors in [1,9] discuss various clustering algorithms, including methods based on the similarity of objects. There are also classic textbooks [2,6,8] that cover the basics and methods of clustering based on similarity and distance. Review materials [3,7] contain numerous clustering methods, emphasizing different approaches, including object similarity. The articles [4,10] explore the reasons for the existence of multiple clustering algorithms and their application depending on the task. The work [5] focuses on clustering algorithms that minimize the maximum distance between clusters, which can be related to the concept of similarity. Spectral clustering [10] uses similarity between objects to form clusters. The proposed sources cover the theoretical foundations and modern approaches to clustering based on the similarity of objects.

The most popular clustering algorithm is the K-means algorithm, which is used to divide into k number of clusters. This algorithm works in the following steps: Choosing a random number of centers (centroids); Sorting each object to the nearest center; Calculation of new centers for each cluster; Steps 2 and 3 are repeated until the centers remain unchanged [11,12]. The next clustering method is Hierarchical Clustering. It is a method of classifying data sets in a hierarchical tree-like structure. This method is implemented by merging data sets from small to large or from large to small [13]. There are two main methods of hierarchical clustering: Agglomerative method: In this method, each data point starts as a separate cluster and then the two closest clusters are combined. This process continues until all data are combined into one large cluster. Divisive method: In this method, all the data starts as one big cluster and then the clusters are divided into parts. This process continues until each data becomes a separate cluster. Also, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method is a density-based clustering method for data

sets. This method is density-based, that is, it identifies clusters by considering the density of points in the dataset. The basic idea of the DBSCAN method is that clusters are treated as densely populated regions of points, and low-density points are separated as "noise"or outliers. [14].

The main steps of the DBSCAN method are as follows:

1. If the number of points within a given radius (epsilon, ε) around a data point exceeds a given minimum density (minPts), this point is called a kernel point.

2. If a point P is a core point and it is connected to all points within a radius e, then they all belong to the same cluster.

3. The points within radius e of the core points, but which are not themselves core points, are called boundary points.

4. All points that are neither core nor boundary points are classified as sum or exception points.

Disadvantages of clustering methods

Disadvantages of clustering methods depend on various factors, and each method has its own limitations and problems.

1. K-means Clustering. In the K-means method, the number of clusters (k) should be predetermined. An improperly chosen number of k can lead to incorrect clustering results.

Sums and outliers can push the centroids too far, distorting the results.

K-means method can only detect spherical clusters, it has difficulty in detecting clusters of other shapes.

2. Hierarchical Clustering. For large datasets, the computational complexity of hierarchical clustering can be enormous, making it difficult to implement.

The inclusion and exclusion of data greatly affects the result of hierarchical clustering.

In hierarchical clustering, the exact number of clusters cannot be defined in advance.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Choosing the appropriate epsilon (ε) and minPts parameters can be complicated. Incorrectly selected parameters will worsen the results.

DBSCAN does not perform well when the data set has different densities.

For large-sized data, the computational complexity of the DBSCAN method is high.

Understanding these shortcomings will help you choose the right clustering method and analyze the results correctly. By knowing the disadvantages of each method, it becomes possible to choose the most suitable method for the data set and the purpose.

The main importance of clustering methods is that they help to identify and understand the internal structure of data [15,16]. Through these methods, it is possible to identify common features within data sets from different fields and make effective decisions based on them [17,18]. With the help of groups, segments and clusters identified through clustering methods, organizations can clearly and effectively focus their strategies.

2 Statement of the clustering problem

Let's suppose that in an N -dimensional character space $x_i \in X$, $i = \overline{1, M}$, objects are given. That is, $x_i = (x_i^1,\ x_i^2,\ \ldots,\ x_i^N)$, $i = \overline{1, M}$, objects are given in N - dimensional symbol space.

Also, in N - dimensional symbol space $\lambda = \left(\lambda^1, \lambda^2, \ldots, \lambda^N\right)$, boolean vectors are introduced, its components take 0 or 1 value. Using components of Vector $\lambda = \left(\lambda^1, \lambda^2, \ldots, \lambda^N\right)$ indicates which character is participating in the calculation or not.

If $\lambda^j = 1$, j-component participates in calculations, otherwise $\lambda^j = 0$, j- component doesn't participates in calculations.

Description. In N- dimensional symbol space $\lambda = \left(\lambda^1, \lambda^2, \ldots, \lambda^N\right)$ Vector ℓ is called informative, if its values taken from the set $\Lambda^\ell = \left\{ \lambda : \sum_{j=1}^N \lambda^j = \ell\ ,\quad \lambda^j \in \{0, 1\}\ ,\ j = \overline{1, N} \right\}$.

Matter. In the given N - dimensional character space $x_i \in X$, $i = \overline{1, M}$ it is required to form training samples using the general selection. Here formation of $x_{p1},\ x_{p2},\ \ldots,\ x_{pm_p} \in X_p$, $p = \overline{1, r}$ classes is required. Here class m_p $x_{p1}, \ldots, x_{p_{m_p}}$ composed of objects $X = \bigcup_{p=1}^r X_p$.

To solve this clustering problem, the object proximity function is introduced. If the signs of the objects are nominal, then the quantity indicating their similarity is determined by $\rho^j(x_{pi}, x_{pq})$:

$$\rho_{pi}^j(x_{pi}, x_{pq}) = \begin{cases} 1, \text{ if } \left(x_{pi}^j - x_{pq}^j\right) = 0; \\ \\ 0, \qquad\quad \text{otherwise} \end{cases} \tag{1}$$

If the objects are represented by numerical symbols, then their similarity is determined by the measuring quantity $\rho^j(x_{pi}, x_{pq})$:

$$\rho_{pi}^j(x_{pi}, x_{pq}) = \begin{cases} 1, \text{ if } \quad |x_{pi}^j - x_{pq}^j| \le \varepsilon^j \\ \\ 0, \qquad\quad \text{otherwise} \end{cases} \tag{2}$$

where $p = \overline{1, r};\ i \ne q = \overline{1, m_p}; j = \overline{1, N};\ \rho_{pi}(x_{pi}, x_{pq}) = (\rho_{pi}^1(x_{pi}, x_{pq}),\ \rho_{pi}^2(x_{pi}, x_{pq}),\ \ldots, \rho_{pi}^N(x_{pi}, x_{pq}))$.

The ε^j -threshold value corresponding to the j-character in the (2) is calculated as follows for the characters of each class of objects j-:

$$\varepsilon^j = \frac{1}{M-1} \sum_{i=1}^{M-1} \left| x_{pi}^j - x_{pi+1}^j \right|,$$

where $j = \overline{1, N}; p = \overline{1, r}; i = \overline{1, M-1};$.

3 The problem of determining the degree of similarity of class objects

Let's suppose that , $\varkappa(x_{pi}, X_p) = \frac{1}{m_p - 1} \sum_{k=1}^{m_p - 1} (\rho_{pi}(x_{pi}, x_{pk}),\ \lambda)$ size x_{pi}, the average degree of similarity of the object X_p with the rest of the objects of the class calculated with respect

to the λ –vector. Likewise, $\varkappa(x_{pi}, X_q) = \frac{1}{m_q} \sum_{k=1}^{m_q} (\rho_{pi}(x_{pi}, x_{qk}),\ \lambda)$ size $x_{pi,}$ calculation of the average degree of similarity of the object with all objects of the X_q class against the λ –vector.

The degree of similarity of the object is defined in percentage by $\nu(x_{pi}, X_p)$ and $\nu(x_{pi}, X_q)$ and they are calculated as follows. In this case, the benefit of the object $\overline{x}$ to each class is calculated as follows.

$$\nu(x_{pi}, X_p) = \frac{\varkappa(x_{pi}, X_p) * 100\%}{N} \tag{3}$$

Equation (3) above the calculation of the value of the object by all x_{pi} objects of the p-class in percentage. Here $p=\overline{1,r}$; $i=\overline{1,m_p}$;.

$$\nu(X_p) = \frac{1}{m_p} \sum_{i=1}^{m_p} \nu(x_{pi}, X_p) = \frac{\frac{1}{m_p} \sum_{i=1}^{m_p} \varkappa(x_{pi}, X_p) * 100\%}{N} \tag{4}$$

Equation (3) finds the average similarity percentage of one object, and formula (4) finds the total average similarity percentage of the objects in s inf, that is, all the objects of the class evaluate each other. Here $p=\overline{1,r}$; $i=\overline{1,m_p}$;.

$$\nu(X_p \ominus \overline{x}) = \frac{1}{m_p - 1} \sum_{i=1}^{m_p - 1} \nu(x_{pi}, X_p) = \frac{\frac{1}{m_p - 1} \sum_{i=1}^{m_p - 1} \varkappa(x_{pi},\ X_p) * 100\%}{N} \tag{5}$$

Equation (5) shows calculation of the total average similarity percentage of the objects in the class after the object $\overline{x}$ of the class X_p is removed from the class, that is, the object does not participate in the evaluation of the remaining objects of the class. As a result, we find the share of the object $\overline{x}$ added to the class X_p by $\nu(X_p) - \nu(X_p \ominus \overline{x})$. Also we determine the position of the object $\overline{x}$ in the organization of its class. Here the operation $X_p \ominus \overline{x}$ means that the object $\overline{x}$ is extracted from the p -class and $p=\overline{1,r}$; $i=\overline{1,m_p-1}$;

$$\nu(x_{qi}, X_q) = \frac{\varkappa(x_{qi},\ \ X_q) * 100\%}{N} \tag{6}$$

Equation (6) given above is the calculation of the percentage of the object x_{qi} by all objects of the q-class. Here $q=\overline{1,r}$; $i=\overline{1,m_q}$; $p \neq q$;

$$\nu(X_q) = \frac{1}{m_q} \sum_{i=1}^{m_q} \nu(x_{qi}, X_q) = \frac{\frac{1}{m_q} \sum_{i=1}^{m_q} \varkappa(x_{qi},\ \ X_q) * 100\%}{N} \tag{7}$$

(7) is used to find the total average similarity percentage of objects in q-class, that is, all the objects of the class evaluate each other. Here $q=\overline{1,r}$; $i=\overline{1,m_q}$; $p \neq q$;

$$\nu(X_q \oplus \overline{x}) = \frac{1}{m_q + 1} \sum_{i=1}^{m_q + 1} \nu\left(x_{qi}, X_q\right) = \frac{\frac{1}{m_q + 1} \sum_{i=1}^{m_q + 1} \varkappa(x_{qi}, X_q) * 100\%}{N} \tag{8}$$

Equation (8) shows the calculation of the total average similarity percentage of the objects in the next class after the object $\overline{x}$ is added to the class X_q. As a result we find by $\nu(X_q \oplus \overline{x}) -$

$\nu(X_q)$ the share of the object $\overline{x}$ that X_q adds to the class. Through this, the share of the object $\overline{x}$ added to all classes is found, and the object is transferred to the class with the largest share. Where the operation $X_q \oplus \overline{x}$ object is added to the q-class. Where $q=\overline{1,r}$; $i=\overline{1,m_q}$; $p \neq q$;

The class to which the evaluated object is transferred in the **first step** for $\overline{x}$ is made as a result of solving the following maximization issue (9).

$$\max_{p,i,q} \{\nu(X_p) \ ,\nu(X_q)\} \tag{9}$$

If
$\max_{p,\mathrm{i},q} \{\nu(X_p) \ ,\nu(X_q)\} = \nu(X_p)$
is left in the $\overline{x} \in X_p$ class, otherwise, ie, if

$$\max_{p,\mathrm{i},q} \{\nu(X_p) \ ,\nu(X_q)\} = \nu(X_q)$$

will be transferred to $\overline{x} \in X_q$ class. If their values are equal, then the object is left in its class. Where $p \neq q = 1,r$; $i = \overline{1,m_p}$. So, on the basis of the proposed algorithm, first, $\nu(X_p)$ and $\nu(X_q)$ values are calculated for arbitrary p,i,q then the maximization problem is solved and the $\overline{x}$ object is transferred to a certain class. Of course, the values of the quantities $\nu(X_p)$ and $\nu(X_q)$ are remembered.

In the second stage $\overline{x}$ – object is evaluated and executed by solving the following maximization problem.

$\max_{p,i,q} \{\nu(X_p) - \nu(X_p \ominus \overline{x}), \nu(X_q \oplus \overline{x}) - \nu(X_q)\}$

If

$\max_{p,i,q} \{\nu(X_p) - \nu(X_p \ominus \overline{x}), \nu(X_q \oplus \overline{x}) - \nu(X_q)\} = \nu(X_p) - \nu(X_p \ominus \overline{x}) > 0,$

it remains at $\overline{x} \in X_p$ and we take it as $\overline{x} = x_{pi+1}$ and the second phase starts again. Otherwise if $\max_{p,i,q} \{\nu(X_p) - \nu(X_p \ominus \overline{x}), \nu(X_q \oplus \overline{x}) - \nu(X_q)\} = \nu(X_p) - \nu(X_p \ominus \overline{x}) \leq 0$, then the class objects remain in place and the second phase starts again.

If it is

$\max_{p,i,q} \{\nu(X_p) - \nu(X_p \ominus \overline{x}), \nu(X_q \oplus \overline{x}) - \nu(X_q)\} = \nu(X_q \oplus \overline{x}) - \nu(X_q) > 0$

will be transferred to $\overline{x} \in X_q$ and will be taken as a $\overline{x} = x_{pi+1}$ and the second phase starts again. Otherwise if will be
$\max_{p,i,q} \{\nu(X_p) - \nu(X_p \ominus \overline{x}), \nu(X_q \oplus \overline{x}) - \nu(X_q)\} = \nu(X_q \oplus \overline{x}) - \nu(X_q) \leq 0$, then the class objects remain in place and the second phase starts again.

The proposed procedure is performed several times for all p,i,q. The procedure also terminates when the resulting class objects stop changing. So, the problem of clustering is solved.

4 Algorithm of clustering based on similarity of objects and analysis of practical results

This research examines clustering by tabulating 1,116 different types of blood pressure medications. One nominal character and nine quantitative characters are studied. They are partially given in the table below.

Table 1

Information about drugs that affect blood pressure (antihypertensive)											
№	The name of the drug	Percentage of active substance in mg	Melting point %	Color	Form of production	Blood concentration (?max)	Distribution density	Technological indicators of prepared compositions			
								Spreadability, 10^{-3} kg/s	Spreading density, kg/m^3	Metabolism in the liver %	Decomposition
1	x_1	0,10	96	yellow	10	64	3,50	4,51	20,72	90	7,41
2	x_2	0,10	85	white	20	78	4,20	7,85	14,36	75	8,56
3	x_3	0,50	88	pale yellow	25	20	4	3,87	25,48	95	5,69
4	x_4	0,03	89	white	50	72	0,20	16,87	15,64	75	7,96
5	x_5	0,20	91	white	15	60	3,20	19,20	25,63	85	5,98
6	x_6	0,50	93	white	25	57	3,90	4,89	48,65	89	6,85
....											
1113	x_{1113}	0,5	76	white	25	86	4,01	7,41	13,25	78	9,65
1114	x_{1114}	0,25	74	gray	14	93	4,5	3,41	10,26	60	12,48
1115	x_{1115}	0,5	72	white	28	94	3,6	4,85	12,48	85	9,85

The following algorithm is proposed to solve the clustering problem. The algorithm consists of the following steps:

Step 1. Drug data is preprocessed into $x_i = (x_i^1,\ x_i^2,\ \ldots,\ x_i^N) \in X,\ i = \overline{1, M}$. In this case, missing data are filled in, anomalous data are replaced by mean values, and based on quantitative symbols $\varepsilon^j = \frac{1}{M-1} \sum_{i=1}^{M-1} \left| x_{pi}^j - x_{pi+1}^j \right|$,

Will be rationing. Where $j=\overline{1,N}; p=\overline{1,r}; i=\overline{1,M-1};$

Step 2. Subjects of training sample are included in the database. The initial database is formed on the intersection of all X_p, $p = \overline{1, r}$ class objects;

Step 3. The magnitude indicating the similarity of objects in the space of nominal symbols and the magnitude indicating the similarity of objects in the space of numerical symbols used in determining the contribution of objects of the class X_p to the formation of their class are calculated using (1) and (2);

Step 4. In the selection symbols, the $i-$object of the optional $p-$class is the remaining place in the set of m_p-1 objects of this class $p-$ all parameters of the vector $\rho_{pi}\left(x_{pi}, x_{pq}\right)$ are calculated for all $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N};$. That is $\rho_{pi}\left(x_{pi}, x_{pq}\right) = \left(\rho_{pi}^1\left(x_{pi}, x_{pq}\right), \rho_{pi}^2\left(x_{pi}, x_{pq}\right)\right), \ldots, \rho_{pi}^N\left(x_{pi}, x_{pq}\right)$ vector symbols $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N};$

For nominal characters

$$\rho_{pi}^j\left(x_{pi}, x_{pq}\right) = \begin{cases} 1, \text{ if } \left(x_{pi}^j - x_{pq}^j\right) = 0; \\ \\ 0, \qquad\quad \text{otherwise} \end{cases}$$

and numeric characters

$$\rho_{pi}^{j}(x_{pi}, x_{pq}) = \begin{cases} 1, \text{ if } & |x_{pi}^{j} - x_{pq}^{j}| \le \varepsilon^{j} \\ 0, & \text{otherwise} \end{cases}$$

calculated based on the above equations;

Step 5. The position of the i - object in the optional p -class in the rest of the set of $m_p - 1$ objects of this class is estimated as follows:

$$\nu(X_p) = \frac{1}{m_p}\sum_{i=1}^{m_p}\nu(x_{pi}, X_p) = \frac{\frac{1}{m_p}\sum_{i=1}^{m_p}\varkappa(x_{pi}, X_p) * 100\%}{N}$$

Step 6. In selection symbols, an arbitrary p-class is the benefit that the i-object brings to this class. In this case, the degree of similarity of the objects in the class after the new object is added to the class

$$\nu(X_q \oplus \overline{x}) = \frac{1}{m_q + 1}\sum_{i=1}^{m_q+1}\nu\left(x_{qi}, X_q\right) = \frac{\frac{1}{m_q+1}\sum_{i=1}^{m_q+1}\varkappa(x_{qi}, X_q) * 100\%}{N}$$

the value before a new object arrives is subtracted,

$$\nu(X_q) = \frac{1}{m_q}\sum_{i=1}^{m_q}\nu(x_{qi}, X_q) = \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\varkappa(x_{qi}, \quad X_q) * 100\%}{N}$$

$\nu(X_q \oplus \overline{x}) - \nu(X_q)$ in this way, an object is transferred to the class with the largest value of the benefits of all classes relative to the new object. If the benefits are equal, they are left in their own class. If the upper values are equal in two classes other than the p-class, then this object is moved to the smaller q.

Based on the proposed theoretical research and algorithm, we solve the problems described above. The experimental results of the clustering algorithm based on the similarity of objects are shown in Table 2. The path elements of the table represent the results of each class at each step after the procedure is executed. The information in column 1 in the table below represents the designation of classes. In this case, the initial training sample is divided into classes $X_1, X_2, X_3, \ldots, X_{10}$. In the initial selection column, the degree of mutual similarity of the objects of the classes and the number of objects at the time of creation of these classes are given. Classification of 1116 drugs was performed in sequence.

The data in the training column 1 is the result of clustering the objects of classes $X_1, X_2, X_3, \ldots, X_{10}$ based on the above algorithm, and we can see that in step 1 itself, the degree of similarity of the objects of the initial classes is increased. The data in training column 1 is used as the initial data for the algorithm in the formation of the data in training column 2. That is, it is performed by rechecking the objects of classes $X_1, X_2, X_3, \ldots, X_{10}$ whose content is newly created. In this, the degree of mutual similarity of objects in some classes and the increase in the number of objects were achieved.

The data in training column 2 is used as the initial data for the algorithm in the formation of the data in training column 3. That is, it is performed by rechecking the objects of classes $X_1, X_2, X_3, \ldots, X_{10}$ whose content is newly created. In this case, the degree of mutual

Table 2: Clustering results based on object similarity clustering algorithm and degree of similarity of clas

№	Initial selection		1-teaching		2-teaching		3-teaching		4-teaching		5-teaching		6-teaching		7-teaching		8-teaching		9-teaching		10-teaching		
	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	Similarity	Number of objects	In his own class
X_1	51.2	271	63.8	56	63.9	188	61.9	276	62.2	254	63.0	242	62.8	222	62.2	210	62.5	198	62.4	204	62.2	199	**65**
X_2	50.8	187	59.2	73	60.1	128	61.2	140	61.6	160	62.5	201	63.5	213	62.6	255	63.1	282	62.4	300	62.0	282	**62**
X_3	49.7	163	60.5	20	59.7	32	60.4	22	67.7	13	66.4	20	65.5	22	67.2	20	69.0	19	67.9	20	65.2	28	**4**
X_4	53.3	107	57.3	180	57.5	112	59.6	51	60.9	47	58.9	71	60.0	54	61.3	60	59.3	71	59.3	69	60.9	64	**9**
X_5	51.1	72	58.1	230	60.5	125	61.9	98	60.0	116	60.2	114	60.3	107	60.0	117	59.2	109	62.0	101	62.3	108	**15**
X_6	54.8	62	66.3	39	62.5	105	62.8	129	62.5	123	62.2	113	62.6	90	61.4	74	64.4	53	64.4	54	64.8	56	**8**
X_7	65.2	7	57.5	62	60.8	62	60.9	72	60.4	68	64.3	53	60.6	107	61.0	86	61.5	104	61.2	109	62.0	137	**2**
X_8	53.4	49	56.3	261	56.9	279	57.1	253	56.9	213	59.4	158	58.3	176	58.3	169	57.8	147	58.7	120	59.6	107	**11**
X_9	51.3	142	52.6	87	56.2	46	58.3	44	59.2	73	58.2	94	59.2	84	58.3	83	59.3	88	58.7	86	58.8	85	**45**
X_{10}	52.9	56	58.9	108	62.5	39	66.8	31	64.1	49	62.6	50	63.8	41	64.3	42	64.0	45	63.3	53	61.8	50	**2**

similarity of objects in some classes is more than 60% and the increase in the number of objects has been achieved.

The data presented in the last column of the table represent the remaining part of objects of classes $X_1, X_2, X_3, \ldots, X_{10}$ in their class. In this case, the original positions of the objects of the classes are preserved, and after the algorithm is completed, it is checked that each object is in the original class. All other columns in the table are also filled in by the result of clustering by the algorithm proposed above.

The data in Table 2 means that initially the degree of similarity of classes was **53.37** percent on average, and after the algorithm was executed once, the degree of similarity of classes was on average **59.05** percent. At the end of the calculations, the degree of similarity of the classes was **61.96** percent on average (Fig. 1).

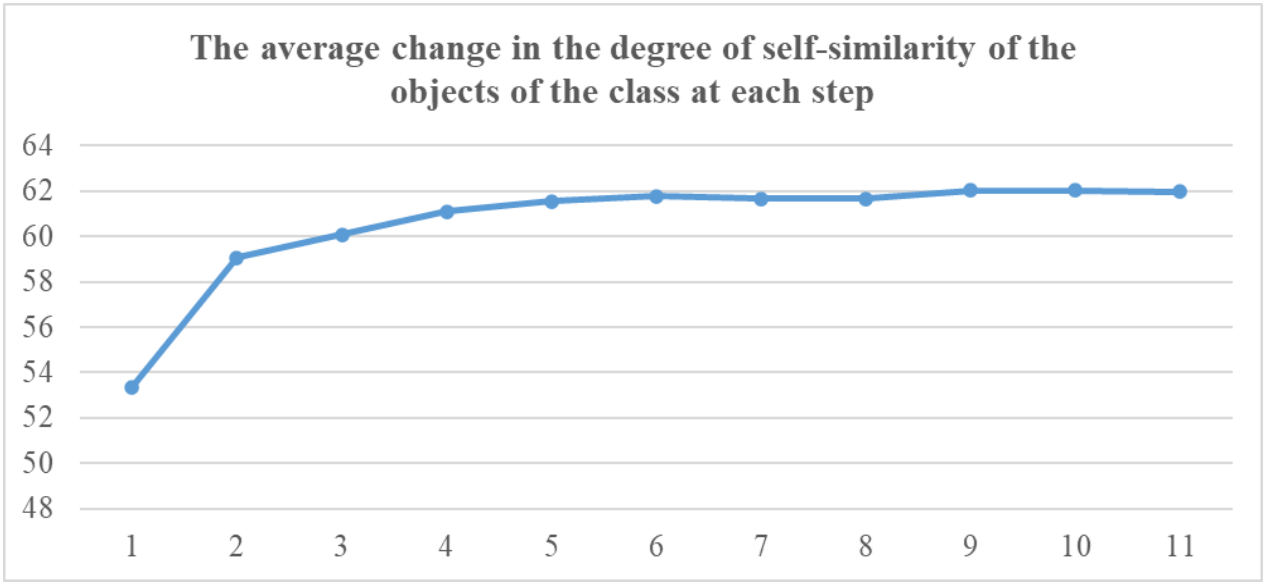

Figure 1: Average change in degrees of similarity between class objects

Based on the data in Table 2, the graph of the change in the degree of similarity of objects of the initial classes in the training sample is presented in Fig.2.

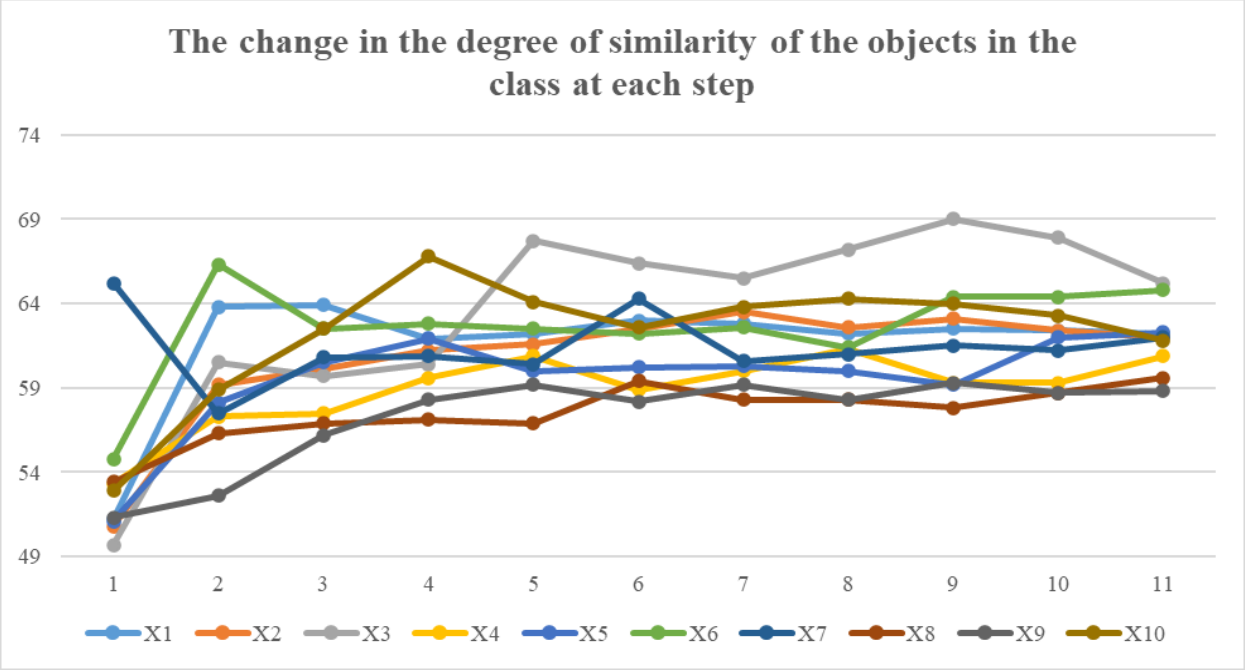

Figure 2: Change of similarity levels of class objects

Based on the results of the clustering algorithm, a graph of the degree of mutual similarity of the objects of the initial classes in the training sample is presented Fig. 3.

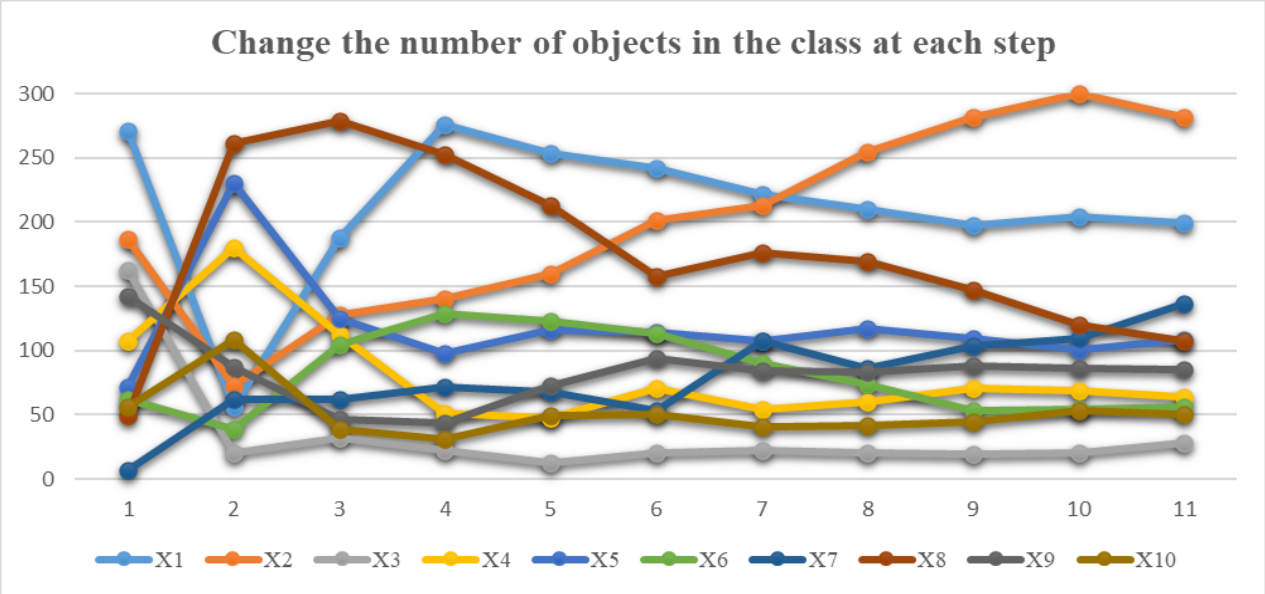

Figure 3: Change in the number of class objects

5 Conclusion

In conclusion, the article proposes a new approach and algorithm for solving the problem of clustering through the space of characters characterizing objects of the training sample class. The primary researched data in the article are different types of drugs that affect blood

pressure taken from real life, and as a result of this research, it was required to solve the problem of clustering by evaluating the similarities between the objects of each class based on the criterion of evaluating the contribution of the objects to the formation of their own class.

Classification, clustering, pattern recognition, problem solving offered by clustering in preprocessing for cases where data symbols are represented by nominal and numerical symbols in intellectual data analysis is a very important research.

We believe that the scientific research proposed in this direction, the method, algorithm and software proposed to solve the problems will not leave the reader indifferent, and we are sure that these studies will find their place in making our lives prosperous and our problems easier.

References

[1] Aggarwal, C. C., & Reddy, C. K. (Eds.). (2013). Data Clustering: Algorithms and Applications. CRC Press.

[2] Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall.

[3] Xu, R., & Wunsch, D. (2005). Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678.

[4] Estivill-Castro, V. (2002). Why So Many Clustering Algorithms: A Position Paper. ACM SIGKDD Explorations Newsletter, 4(1), 65-75.

[5] Gonzalez, T. F. (1985). Clustering to Minimize the Maximum Intercluster Distance. Theoretical Computer Science, 38, 293-306.

[6] Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern Classification. John Wiley & Sons.

[7] Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

[8] Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

[9] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

[10] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On Spectral Clustering: Analysis and an Algorithm. Advances in Neural Information Processing Systems, 14.

[11] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.

[12] Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

[13] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[14] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 226-231).

[15] Nishanov A. Ruzibaev O. Tran N. Modification of decision rules 'ball Apolonia' the problem of classification // 2016 International Conference on Information Science and Communications Technologies, ICISCT 2016 (2016), doi: 10.1109/ICISCT.2016.7777382.

[16] Nishanov A., Saidrasulov Sh., Babadjanov E. Analysis of Methodology Of Rating Evaluation Of Digital Economy And E-Government Development In Uzbekistan // International Journal Of Early Childhood Special Education, Volume:14 Issue:2 2022, p. 2447-2452, doi: 10.9756/INT-JECSE/V14I2.230.

[17] Nishanov A., Ruzibaev O., Chedjou J. C., Kyamakya K., Abhiram, Kolli, De Silva, Djurayev G., Khasanova M. Algorithm for the selection of informative symptoms in the classification of medical data // Developments Of Artificial Intelligence Technologies In Computation And Robotics, Volume:12, 2020, p. 647-658, doi:10.1142/9789811223334_0078

[18] Nishanov AKh, Turakulov AKh, Turakhanov KhV. Reshaiushchee pravilo dlia klassifikatsii patologii zritel'noĭ sistemy [A decisive rule in classifying diseases of the visual system]. Med Tekh. 1999 Jul-Aug; 4:16-8. Russian. PMID: 10464756.

Information about authors:

Nishanov Akhram Khasanovich (corresponding author) – DSc, professor of the Faculty of Software engineering of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi (Tashkent, Uzbekistan, email: nishanov_ akram@mail.ru);

Akbarova Marguba Khamidovna – PhD, associate professor of the Faculty of Software engineering of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi (Tashkent, Uzbekistan, email: margubaakbarova66@gmail.com);

Tursunov Alisher Tulkunovich – Senior lecturer of the Department of Physics, Mathematics and Information Technologies of Tashkent Pharmaceutical Institute (Tashkent, Uzbekistan, email: tursunovr484za@gmail.com);

Ollamberganov Fayzulla Farxod o'g'li – PhD student Department of System and applied programming of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi (Tashkent, Uzbekistan, email: fayzulla0804@gmail.com);

Rashidova Dilfuza Elmurodovna – Teacher of the Department of Information Technologies of the Samarkand Institute of Economics and Service (Samarkand, Uzbekistan, email: dilfuzarashidova23@gmail.com).

Информация об авторах: *Нишанов Ахрам Хасанович (ответственный автор) – доктор технических наук, профессор факультета программной инженерии Ташкентского*

университета информационных технологий имени Мухаммада Ал-Хоразмий (Ташкент, Узбекистан, почта: nishanov_ akram@mail.ru);

Акбарова Маргуба Хамидовна – ф-м.ф.д., доцент факультета программная инженерии Ташкентского университета информационных технологий имени Мухаммада Ал-Хоразмий (Ташкент, Узбекистан, почта: margubaakbarova66@gmail.com);

Турсунов Алишер Тулкунович – Старший преподаватель кафедры физики, математики и информационных технологий Ташкентского фармацевтического института (Ташкент, Узбекистан, почта: tursunovr484za@gmail.com);

Олламберганов Файзулла Фарход угли – докторант кафедры системного и прикладного программирования Ташкентского университета информационных технологий имени Мухаммада Ал-Хоразмий (Ташкент, Узбекистан, почта: fayzulla0804@gmail.com);

Рашидова Дильфуза Элмуродовна – преподаватель кафедры информационных технологий Самаркандского института экономики и сервиса (Самаканд, Узбекистан, почта: dilfuzarashidova23@gmail.com).

Received: August 29, 2024
Accepted: September 16, 2024