Chemical Technology, Control and Management

Manuscript 1586

CLUSTERING ALGORITHM BASED ON SIMILARITY OF OBJECTS

Akhram Khasanovich Nishanov

Alisher Tulkunovich Tursunov

Fayzullo Farhod o'g'li Ollamberganov

ISSN 1815-4840, E-ISSN 2181-1105
Himičeskaâ tehnologiâ. Kontrol′ i upravlenie
CHEMICAL TECHNOLOGY. CONTROL AND MANAGEMENT
2024, №4 (118) pp.11-21
International scientific and technical journal
journal homepage: https://ijctcm.researchcommons.org/journal/

Article history: Received 30 May 2024; Received in revised form 01 July 2024; Accepted 30 August 2024; Available online 28 September 2024

Since 2005

UDC 004.931

CLUSTERING ALGORITHM BASED ON SIMILARITY OF OBJECTS (АЛГОРИТМ КЛАСТЕРИЗАЦИИ, ОСНОВАННЫЙ НА СХОДСТВЕ ОБЪЕКТОВ)

Nishanov Akhram Khasanovich[1], Tursunov Alisher Tulkunovich[2], Ollamberganov Fayzulla Farhod o'g'li[3]

[1,3]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi.
Address: Amir Temur Ave. 108, 100084, Tashkent, Republic of Uzbekistan.
E-mail: [1]nishanov_akram@mail.ru, Phone: +998935992922;
[2]Tashkent Pharmaceutical Institute. Address: Oybek Ave. 45, 100045, Tashkent city, Republic of Uzbekistan.
E-mail: tursunovr484za@gmail.com, Phone: +998884115353.

Abstract: *The article discusses the problem of drug clustering. Initially, k classes are randomly formed and the resulting training sample is preprocessed, then the similarity between objects of each class is assessed based on the proximity function and the criterion for assessing the contribution of objects to the formation of their own class. It is usually expressed as a percentage and represents 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 remains in the class that contributed the most. This process is repeated several times in a row for all objects of the class. The process stops when the location of objects does not change and the degree of similarity exceeds the required percentage. As a result, the necessary clusters are formed.*

Keywords: *Clustering problem, proximity function, degree of similarity of objects, contribution of an object to a class, agglomerative method, partitioning method, DBSCAN.*

Annotatsiya*: Maqolada dori vositalarini klasterlash masalasi tadqiq etilgan. Dastlab, ixtiyoriy ravishda k-ta sinf shakllantiriladi va hosil bo'lgan o'quv tanlanmaga dastlabki ishlov beriladi, so'ngra yaqinlik funksiyasi va obyektlarni o'z sinfini shakllanishiga qo'shgan hissasini baholash mezoni asosida har bir sinf obyektlari o'rtasidagi o'xshashliklar baholanadi. Odatda, bu foizda bo'lib, har bir sinf obyektlarining o'zaro o'xshashlik darajasi hisoblanadi. Algoritmning keyingi qadamlarida dastlab, birinchi sinfdan bitta obyekt olinadi va barcha k-ta sinflarga uni qo'shib ushbu obyektni shu sinfga qo'shgan hissasi o'lchanadi. Qaysi sinfga hissasi ko'p tegsa obyekt o'sha sinfda qoldiriladi. Bu jarayon barcha sinf obyektlari uchun ketma-ket to'liq ravishda bir necha marta qaytariladi. Obyektlar joyi o'zgarmay qolganda va o'xshashlik darajasi kerakli foizdan oshganda jarayon to'xtatiladi. Natijada, talab qilingan klasterlar shakllanadi.*

Tayanch so'zlar. *klasterlash masalasi, yaqinlik funksiyasi, obyektlarning o'xshashlik darajasi, obyektni sinfga qo'shgan hissasi, aglomerativ (qo'shish) usuli, diviziv (ajratish) usuli, DBSCAN.*

Аннотация*: В статье рассматривается проблема кластеризации лекарственных средств. Первоначально произвольно формируются k классов и предварительно обрабатывается полученная обучающая выборка, затем на основе функции близости и критерия оценки вклада объектов в формирование собственного класса оценивается сходство между объектами каждого класса. Обычно оно 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 остается в классе, внёсшем 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 необходимые кластеры .*

Ключевые слова: *задача кластеризации, функция близости, степень сходства объектов, вклад объекта в класс, агломеративный метод, метод разделения, DBSCAN.*

Введение

Алгоритмы кластеризации, основанные на сходстве объектов, используются для поиска сходства между данными и разделения их на кластеры. Эти алгоритмы работают на основе критериев сходства или расстояния внутри данных. Основная философия этих алгоритмов заключается в том, что объекты внутри каждого кластера схожи, но объекты между разными кластерами различны.

Самым популярным алгоритмом кластеризации является алгоритм K-means, который используется для разделения на k кластеров. Этот алгоритм работает в несколько этапов:

1. Выбор случайного количества центров (центроидов);
2. Сортировка каждого объекта до ближайшего центра;
3. Расчет новых центров для каждого кластера;
4. Шаги 2 и 3 повторяются до тех пор, пока центры не останутся неизменными [1,2].

Следующий метод кластеризации — иерархическая кластеризация (Hierarchical Clustering). Это метод классификации наборов данных в иерархическую древовидную структуру. Этот метод реализуется путем объединения наборов данных от малого к большому или от большого к малому.

Существует два основных метода иерархической кластеризации:

Агломеративный метод . В этом методе каждая точка данных начинается как отдельный кластер, а затем два ближайших кластера объединяются. Этот процесс продолжается до тех пор, пока все данные не будут объединены в один большой кластер.

Метод разделения. В этом методе все данные начинаются как один большой кластер, а затем кластеры делятся на части. Этот процесс продолжается до тех пор, пока каждые данные не станут отдельным кластером [3].

Кроме того, метод DBSCAN (пространственная кластеризация приложений с шумом на основе плотности) — это метод кластеризации наборов данных на основе плотности. Этот метод основан на плотности, то есть он идентифицирует кластеры, учитывая плотность точек в наборе данных. Основная идея метода DBSCAN заключается в том, что кластеры рассматриваются как густонаселённые области точек, а точки с низкой плотностью выделяются как «шумы» или выбросы [4].

Основные этапы метода DBSCAN следующие:

1. Если количество точек в пределах заданного радиуса (epsilon, ε) вокруг точки данных превышает заданную минимальную плотность (minPts), эта точка называется точкой ядра.

2. Если точка P является центральной точкой и соединена со всеми точками радиуса e, то все они принадлежат одному кластеру.

3. Точки внутри радиуса ε основных точек, но которые сами по себе не являются основными точками, называются граничными точками.

4. Все точки, которые не являются ни основными, ни граничными точками, классифицируются как точки суммы или точки исключения.

Недостатки методов кластеризации зависят от различных факторов, и каждый метод имеет свои ограничения и проблемы.

1. K-means Clustering. (Кластеризация методом k-средних) В методе K-means количество кластеров (k) должно быть заранее определено. Неправильно выбранное число k может привести к неверным результатам кластеризации.

Суммы и выбросы могут слишком сильно сдвинуть центроиды, исказив результаты.

Метод K-means может обнаруживать только сферические кластеры, но с трудом обнаруживает кластеры других форм.

2. Hierarchical Clustering. (Иерархическая кластеризация) Для больших наборов данных вычислительная сложность иерархической кластеризации может быть огромной, что затрудняет ее реализацию.

Включение и исключение данных сильно влияет на результат иерархической кластеризации.

При иерархической кластеризации точное количество кластеров не может быть определено заранее.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Выбор подходящих параметров эпсилон (ε) и minPts может оказаться сложной задачей. Неправильно выбранные параметры ухудшат результаты.

DBSCAN не работает должным образом, если набор данных имеет разную плотность.

Для данных большого размера вычислительная сложность метода DBSCAN высока.

Понимание этих недостатков поможет вам выбрать правильный метод кластеризации и правильно проанализировать результаты. Зная недостатки каждого метода, становится возможным выбрать наиболее подходящий метод для набора данных и цели.

Основное назначение методов кластеризации заключается в том, что они помогают идентифицировать и понять внутреннюю структуру данных [5,6]. С помощью этих методов можно выявить общие черты в наборах данных из разных областей и принимать на их основе эффективные решения [7,8]. С помощью групп, сегментов и кластеров, определенных с помощью методов кластеризации, организации могут четко и эффективно сфокусировать свои стратегии.

Методы исследования и полученные результаты

Выражение вопроса кластеризации. Предположим, что в *N*-мерном пространстве знаков заданы объекты $x_i \in X$, $i = \overline{1, M}$,. То – есть, объекты $x_i = (x_i^1, x_i^2, \dots, x_i^N)$, $i = \overline{1, M}$, заданы в *N*-мерном пространстве знаков.

Кроме того, в N - мерном символьном пространстве вводятся векторы bul$\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$, компоненты которых принимают нулевое или единичное значение. С помощью вектора $\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$, использование компонентов которого указывает, какой знак участвует или не участвует в вычислительной работе.

Если $\lambda^j = 1$, то компонент j- участвует в работе вычисления, в противном случае если $\lambda^j = 0$, то компонент j- не участвует в работе вычисления.

Определение. в $N-$мерном пространстве знаков вектор $\lambda = (\lambda^1, \lambda^2, \dots, \lambda^N)$ называется ℓ информативным, если его значения берутся из комплекса $\Lambda^\ell = \left\{ \lambda: \sum_{j=1}^N \lambda^j = \ell\,,\ \lambda^j \in \{0,1\}, j = \overline{1, N} \right\}$.

Вопрос. Требуется сформировать учебные выборки $x_i \in X,$ в заданном N - мерном символьном пространстве, используя из общей выборки $i = \overline{1, M}$. При этом требуется сформировать классы $x_{p1}, x_{p2}, \dots, x_{pm_p} \in X_p, p = \overline{1, r}$. Здесь класс X_p будет $X == \cup_{p=1}^r X_p$ который состоит из m_p объектов $x_{p1}, \dots, x_{p_{m_p}}$.

Чтобы решить эту проблему кластеризации, вводится функция близости между объектами. Если значения объектов являются номинальными, то величина, указывающая на их сходство, определяется путем $\rho^j(x_{pi}, x_{pq})$ и измеряется следующей формулой:

$$\rho_{pi}^j(x_{pi}, x_{pq}) = \begin{cases} 1, & \text{если} \quad (x_{pi}^j - x_{pq}^j) = 0; \\ 0, & \text{то} \end{cases} \tag{1}$$

Если объекты представлены числовыми символами, то величина, измеряющая их сходство, определяется путем $\rho^j(x_{pi}, x_{pq})$ и исчисляется формулой (2)

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

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

Дробное значение $\varepsilon^j -$, соответствующее знаку j в приведенной выше формуле (2), вычисляется для знаков объектов каждого класса следующим образом:

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

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

Вопрос рассмотрения определения степени сходства объектов класса

Предположим, что величина $\varkappa(x_{pi}, X_p) = \frac{1}{m_p-1}\sum_{k=1}^{m_p-1}(\rho_{pi}(x_{pi}, x_{pk}), \lambda)$ - $x_{pi,}$ вычисление степени среднего сходства с остальными объектами класса X_p по отношению к вектрору λ. Точно также величина $\varkappa(x_{pi}, X_q) = \frac{1}{m_q}\sum_{k=1}^{m_q}(\rho_{pi}(x_{pi}, x_{qk}), \lambda)$ - $x_{pi,}$ вычисление степени среднего сходства со всеми объектами класса X_q по отношению к вектрору λ.

Степени сходства объектов в процентах определяются через $\nu(x_{pi}, X_p)$ и $\nu(x_{pi}, X_q)$, и они вычисляются следующим образом. Здесь польза, которую приносит каждому классу объекта $\bar{x}$, вычисляется следующим образом:

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

Вычисление в процентах оценки расстояния между всеми объектами класса p- объекта x_{pi} определяется вышеуказанной формулой (3). Здесь $p = \overline{1,r};\ i = \overline{1,m_p};$.

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

Если найдено среднее сходство одного объекта через формулу (3), то будет найден процент общего среднего сходства объектов в классе путем формулы (4), то есть здесь все объектов класса оценивают друг друга. Здесь $p = \overline{1,r};\ i = \overline{1,m_p};$

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

Вычисление процента общего среднего сходства объектов после вывода из класса объекта $\bar{x}$ класса X_p через формулу (5), то есть объект $\bar{x}$ не участвует при оценке остальных объектов класса. В итоге мы найдем долю вклада объекта $\bar{x}$ в класс X_p через $\nu(X_p) - \nu(X_p \ominus \bar{x})$. Также определим место состава своего класса объекта $\bar{x}$. Здесь действие $X_p \ominus \bar{x}$ означает вывод объекта $\bar{x}$ из класса p- и $p = \overline{1,r};\ i = \overline{1,m_p-1};$

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

Вычисление в процентах оценки всеми объектами класса q объекта x_{qi} определяется по вышеуказанной формуле (6). Здесь $q = \overline{1,r};\ i = \overline{1,m_q};\ p \neq q;$

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

Через формулу (7) мы найдем процент общего среднего сходства объекта класса q-, то есть здесь все объекты класса оценивают друг друга. Здесь $q = \overline{1,r};\ i = \overline{1,m_q};\ p \neq q;$

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

В этом при вычислении процента общего среднего сходства объекта следующего класса после присоединения объекта $\bar{x}$ в класс X_q мы находим долю объекта $\bar{x}$, внесенную в класс X_q через $\nu(X_q \oplus \bar{x}) - \nu(X_q)$. Таким образом найдена доля объекта $\bar{x}$, внесённая во все классы, и

объект переводится в класс, имеющий наибольшую долю. Здесь действие $X_q \oplus \bar{x}$ выражает присоединение объекта $\bar{x}$ в класс q-; $q = \overline{1,r};\ i = \overline{1,m_q};\ p \neq q$;

Для определения вопроса перевода в какой то класс оцениваемого объекта $\bar{x}$ **на первом этапе** осуществляется в результате решения следующего вопроса максимизации:

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

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

то $\bar{x}$ остается в классе $\in X_p$, в противном случае, если

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

то $\bar{x}$ переводится в класс $\in X_q$. Если их значение равно, то объект оставляется в своем классе.Здесь $p \neq q = \overline{1,r};\ i = \overline{1,m_p}$. Итак, первоначально на основе предлагаемого алгоритма высчисляются величины $\nu(X_p)$ и $\nu(X_q)$ для произвольных p, i, q , затем решается вопрос максимизации, и объект $\bar{x}$ переводится в какой-либо класс. Обязательно значения величин $\nu(X_p)$ и $\nu(X_q)$ сохраняются.

На втором этапе объект $\bar{x}$ – оценивается и осуществляется в результате решения следующего вопроса максимизации.

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

Если

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

то $\bar{x}$ остается в $\in X_p$и принимается $\bar{x} = x_{pi+1}$, и второй этап начинается заново. В противном случае если

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

то объекты класса остаются на своих местах и второй этап начинается заново.

Если

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

то $\bar{x}$ переводится в $\in X_q$ и принимается как $\bar{x} = x_{pi+1}$ и второй этап начинается заново. В пртивном случае если

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

то объекты класса остаются на своих местах и второй этап начинается заново.

Предлагаемая процедура осуществляется несколько раз для всех p, i, q. Когда изменение объектов образовавшихся классов останавливается, процедура также заканчивается. Значит, вопрос кластеризации будет решен [9,10,11].

Алгоритм кластеризации на основе сходства объектов и анализ практических результатов. Была изучена кластеризация, результаты которой были сведены в таблицы 1116 различных препаратов от артериального давления. Изучены один номинальный признак и девять количественных признаков. Частично они представлены в таблице ниже.

Таблица 1

Информация о препаратах, влияющих на артериальное давление (гипотензивных)

Нет	Название препарата	Процент активного	Эрозия %	Цвет	Форма	Концентрация в	Плотность	Технологические показатели приготовленных составов			
								Растекаемость, 10^{-3} кг/с	Плотность укладки, кг/м^3	Метаболизм в печени, %	Разложение
1	x_1	0,10	96	желтый	10	64	3.50	4.51	20.72	90	7.41
2	x_2	0,10	85	белый	20	78	4.20	7,85	14.36	75	8.56
3	x_3	0,50	88	бледно-желтый	25	20	4	3,87	25.48	95	5,69
4	x_4	0,03	89	белый	50	72	0,20	16.87	15,64	75	7,96
5	x_5	0,20	91	белый	15	60	3.20	19,20	25,63	85	5,98
6	x_6	0,50	93	белый	25	57	3,90	4,89	48,65	89	6,85
..... …											
1113	x_{1113}	0,5	76	белый	25	86	4.01	7.41	13.25	78	9,65
1114	x_{1114}	0,25	74	серый	14	93	4,5	3.41	10.26	60	12.48
1115	x_{1115}	0,5	72	белый	28	94	3.6	4,85	12.48	85	9,85

Для того, чтобы найти решение вопроса кластеризации, предлагается алгоритм, который состоит из следующих пунктов:

*Первый шаг***.** Информация о лекарственных средствах $x_i = (x_i^1, x_i^2, \dots, x_i^N) \in X,\ i = \overline{1, M}$ подвергается первоначальной обработке. Здесь пропущенная информация дополняется, аномальная информация заменяется средним значением и нормируется на основе количественных значений

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

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

*Второй шаг***.** Учебно-выборочные **о**бъекты вводятся в базу данных. Первичная база данных формируется в разрезе всех X_p, объектов класса $p = \overline{1, r}$;

*Третий шаг***.** В пространстве номинальных значений, используемых при определении оценки доли, внесенной в формирование объектов класса X_p, которые они добавили к формированию своего класса величина, указывающая на сходство объектов, и величина, указывающая на сходство объектов в пространстве числовых символов, вычисляются по величинной формуле;

*Четвертый шаг***.** Место занимающее в комплексе объектов 1 в этом классе $m_p -$, где остался объект $i -$ в выборочном классе $p -$ в выборочных знаках, считаются все параметры вектора $\rho_{pi}(x_{pi,}x_{pq})$ для всех $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N};$ То есть, знаки вектора $\rho_{pi}(x_{pi,}x_{pq}) = (\rho_{pi}^1(x_{pi,}x_{pq}), \rho_{pi}^2(x_{pi,}x_{pq}), \dots, \rho_{pi}^N(x_{pi,}x_{pq})$ для номинальных знаков $p = \overline{1, r};\ i \neq q = \overline{1, m_p}; j = \overline{1, N};$

Вычисляются по формуле:

$$\rho_{pi}^j(x_{pi,}x_{pq}) = \begin{cases} 1, & \text{если } (x_{pi}^j - x_{pq}^j) = 0; \\ 0, & \text{то} \end{cases}$$

Числовые значения

$$\rho_{pi}^j(x_{pi,}x_{pq}) = \begin{cases} 1, \text{если } |x_{pi}^j - x_{pq}^j| \leq \varepsilon^j \\ 0, \quad \text{то} \end{cases}$$

Пятый шаг. Место, имеющее в комплексе 1 объекта в этом классе где остался объект i в произвольном p-классе, оценивается в процентах следующим образом и рассчитывается на основе формулы:

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

Шестой шаг. i −объект в любом классе p в выборочных значениях является пользой, которую он привносит в этот класс. В этом случае значение до появления нового объекта вычитается из уровня сходства объектов в следующем классе

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

после добавления нового объекта,

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

$\nu(X_q \oplus \bar{x}) - \nu(X_q)$ таким образом, объект переносится в класс с наибольшим значением преимуществ из всех классов по отношению к новому объекту. если преимущества равны - то оставляется в своем классе. Если более высокие значения равны в двух других классах, а не в p-классе, то этот объект переносится в меньший класс q.

На основе предложенного алгоритма теоретического исследования мы решим вышеописанные проблемы. Результаты кластеризации на основе алгоритма кластеризации, основанного на сходстве объектов, отражены во второй таблице ниже. На основе процедуры, описанной выше, была завершена следующая таблица-2. Элементы таблицы шагов указывают на результаты каждого класса на каждом этапе после выполнения процедуры. Данные в столбце 1, приведенные в таблице ниже, представляют собой обозначения классов. Начальная обучающая выборка в этом случае разделяется на классы $X_1, X_2, X_3, \dots, X_{10}$. В столбце начальной выборки указаны уровни сходства и количество объектов классов, в которые эти классы объединены. При классификации по классам 1116 препаратов были применены в определенной последовательности.

Данные в 1-м обучающем столбце являются результатом кластеризации объектов представленного выше объектов классов $X_1, X_2, X_3, \dots, X_{10}$ на основе алгоритма, и мы можем видеть, что на самом первом шаге 1 достигается увеличение степени взаимного сходства первоначальных объектов классов [12,13].

Таблица 2

Результаты кластеризации и классы объектов на базе алгоритма кластеризации, основанного на сходстве объектов по степени сходства

№	Первоначальный выбор		1- Обучение		2- Обучение		3- Обучение		4- Обучение		5- Обучение		6- Обучение		7- Обучение		8- Обучение		9- Обучение		10- Обучение		В своем классе
	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	Сходство	Количество объектов	
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**

Данные в обучающем столбце 1 используются в качестве исходных данных для алгоритма формирования данных в обучающем столбце 2. То есть это делается путем перепроверки объектов классов $X_1, X_2, X_3, \dots, X_{10}$ контент которых создан заново. При этом достигалась степень взаимного сходства объектов в некоторых классах и увеличение количества объектов.

Данные в обучающем столбце 2 используются в качестве исходных данных для алгоритма при формировании данных в обучающем столбце 3. То есть содержание осуществляется путем перепроверки объектов классов $X_1, X_2, X_3, \dots, X_{10}$. При этом степень взаимного сходства объектов в некоторых классах составляет более 60% и количество объектов увеличилось.

Все остальные столбцы таблицы также заполняются результатом кластеризации по предложенному выше алгоритму.

Данные, представленные в последнем столбце таблицы, представляют остальную часть объектов классов $X_1, X_2, X_3, \dots, X_{10}$ в своем классе. При этом исходные положения объектов классов сохранялись и после завершения работы алгоритма проверялось, что каждый объект находится в исходном классе.[14,15,16,17]

Данные таблицы означают, что изначально степень сходства классов составляла в среднем 53,37 процента, а после однократного выполнения алгоритма степень сходства классов составляла в среднем 59,05 процента. По итогам расчетов степень сходства классов составила в среднем 61,96 процента (рис. 1).

Рис. 1. Среднее изменение степени сходства между объектами класса.

На основе данных приведенной таблицы представлен график изменения уровней сходства объектов начальных классов в образовательной выборке (рис. 2).

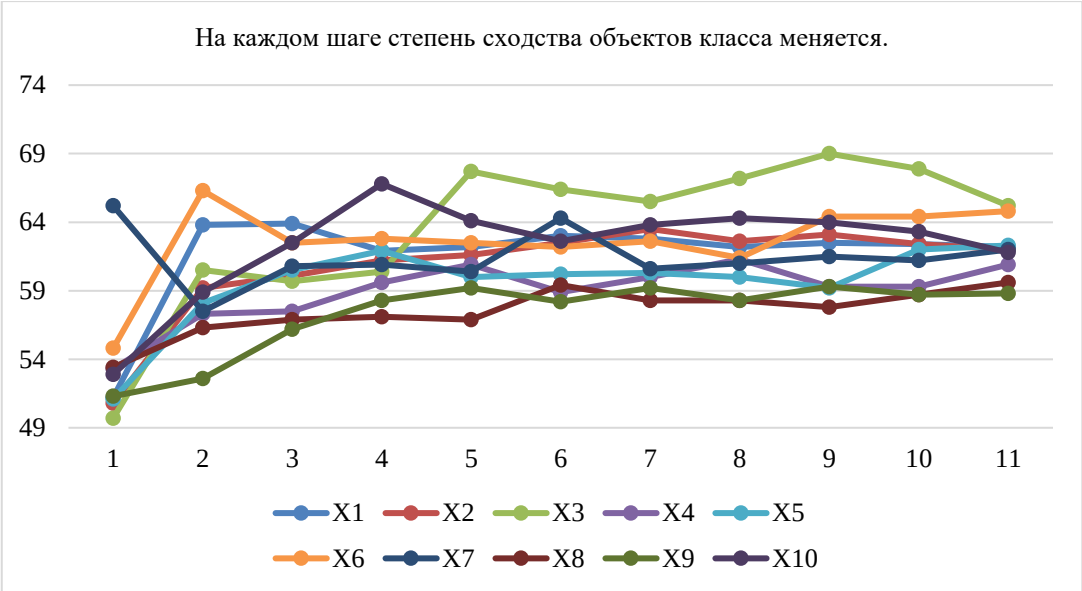

Рис. 2. ***Изменение степени сходства объектов класса.***

По результатам кластеризации на основе алгоритма кластеризации по сходству объектов представлен график степени взаимного сходства объектов первых классов образовательной выборки (рис. 3).

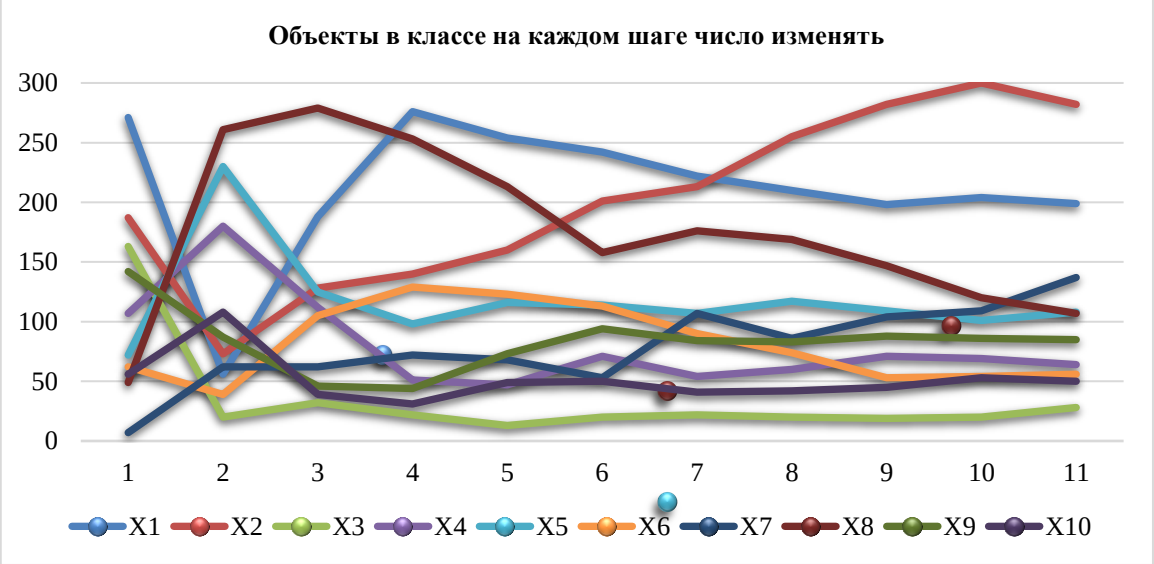

Рис. 3. ***Изменение количества объектов класса.***

Заключение

В статье предлагаются новый подход и алгоритм решения задачи кластеризации через пространство признаков, характеризующих объекты класса образовательного отбора.

Первичной исследуемой информацией в статье являются различные виды препаратов, влияющих на артериальное давление, взятые из реальной жизни, и в результате данного исследования потребовалось решить задачу кластеризации путем оценки сходства между объектами каждого класса на основе критерия оценки вклада объектов в формирование собственного класса.

Классификация, кластеризация, распознавание образов, решение задач, предлагаемых кластеризацией при предварительной обработке для случаев, когда символы данных

представлены именными и числовыми символами при интеллектуальном анализе данных, являются очень важными исследованиями.

Мы думаем, что предлагаемые научные исследования в этом направлении, предлагаемый метод, алгоритм и программные исследования решения поставленных задач не оставят равнодушным читателя, и мы уверены, что эти исследования найдут свое место в том, чтобы сделать нашу жизнь благополучной и решить наши проблемы.

References

1. Witten, I.H., Frank, Ye., Hall, M.A. (2011). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
2. Arthur, D., Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the yeighteenth annual ACM-SIAM symposium on Discrete algorithms*. *Society for Industrial and Applied Mathematics*. 1027-1035.
3. Hastie, T., Tibshirani, R., Friedman, J. (2009). The Yelements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Science & Business Media*.
4. Yester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*. 226-231.
5. 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*. doi: 10.1109/ICISCT.2016.7777382.
6. Nishanov, A., Saidrasulov, Sh., Babadjanov, Ye. (2022). Analysis of Methodology Of Rating Yevaluation Of Digital Yeconomy And Ye-Government Development In Uzbekistan. *International Journal Of Yearly Childhood Special Yeducation*, 14(2), 2447-2452. doi: 10.9756/INT-JECSE/V14I2.230.
7. Nishanov, A., Ruzibaev, O., Chedjou, J.C., Kyamakya, K., Abhiram, Kolli, De Silva, Djurayev, G., Khasanova, M. (2020). Algorithm for the selection of informative symptoms in the classification of medical data. *Developments Of Artificial Intelligence Technologies In Computation And Robotics*. 12, 647-658. doi:10.1142/9789811223334_0078
8. Nishanov, A.Kh., Turakulov, A.Kh., Turakhanov, Kh.V. (1999). Reshaiushchee pravilo dlia klassifikatsii patologii zritel'noĭ sistemy [A decisive rule in classifying diseases of the visual system]. *Med Tekh*. (4):16-8. (in Russian).
9. Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*. 31(8), 651-666. DOI: 10.1016/j.patrec.2009.09.011.
10. Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika*. 32(3), 241-254. DOI: 10.1007/BF02289588
11. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96(34), 226-231.
12. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1), 1-38. DOI: 10.1111/j.2517-6161.1977.tb01600.
13. Ng, A.Y., Jordan, M.I., Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2, 849-856.
14. Huang, X., Yang, Q., Yu, Y. (2008). Towards Context-Aware Search by Learning a Very Large Variable Length Hidden Markov Model from Search Logs. *Proceedings of the 17th ACM Conference on Information and Knowledge Management - CIKM '08*, 1197-1206. DOI: 10.1145/1458082.1458243.
15. Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2), 1-135. DOI: 10.1561/1500000011.
16. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167. DOI: 10.2200/S00416ED1V01Y201204HLT016.
17. Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.