ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 68 за 2022 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Н.Д. Баданина, А.А. Зинченко, В.А. Судаков

Ранжирование объектов на основе нечеткой кластеризации

Рекомендуемая форма библиографической ссылки: Баданина Н.Д., Зинченко А.А., Судаков В.А. Ранжирование объектов на основе нечеткой кластеризации // Препринты ИПМ им. М.В.Келдыша. 2022. № 68. 12 с. https://doi.org/10.20948/prepr-2022-68
https://library.keldysh.ru/preprint.asp?id=2022-68

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Н.Д. Баданина, А.А. Зинченко, В.А. Судаков

Ранжирование объектов на основе нечеткой кластеризации

Москва — 2022

Баданина Н.Д, Зинченко А.А., Судаков В.А.

Ранжирование объектов на основе нечеткой кластеризации

В работе создана методика применения нечеткой логики в задаче разделения объектов на кластеры и их ранжирования, где в качестве признаков указываются их координаты. Исследуются классические подходы машинного обучения для составления моделей кластеризации. Далее они дополнены аппаратом нечетких чисел для получения оценки потенциальной возможности принадлежности объекта кластеру. На основе выбранных подходов было разработано алгоритмическое и программное обеспечение для отнесения объекта к кластерам с выводом функции принадлежности, а также выводом ранга, вычисляемого через дефаззификацию с учетом важности каждого кластера. Полученная модель может быть использована для решения задач выбора и ранжирования объектов с учетом степени уверенности в принадлежности их определенным классам.

Ключевые слова: кластеризация, нечеткая логика, метод k-средних, мягкие вычисления, машинное обучение

Natalya Dmitriyevna Badanina, Aleksandra Andreevna Zinchenko, Vladimir Anatolyevich Sudakov

Ranking of objects based on fuzzy clustering

The paper creates a method for using fuzzy logic in the task of dividing objects into clusters and ranking them, where their coordinates are indicated as features. Classical machine learning approaches for making clustering models are investigated. Further, they are supplemented with the fuzzy numbers apparatus to obtain an estimate of the potential possibility of an object belonging to a cluster. Based on the selected approaches, algorithmic and software were developed for assigning an object to clusters with the derivation of the membership function, as well as the derivation of the rank calculated through defuzzification, taking into account the importance of each cluster. The resulting model can be used to solve the problems of selecting and ranking objects, taking into account the degree of confidence in their belonging to certain classes.

Key words: Clustering, fuzzy logic, k-means method, soft computing, machine learning

Исследование выполнено при финансовой поддержке РФФИ и CNPq (Бразилия), Фонда содействия инновациям (Россия), DBT, DST (Индия), MOST, NSFC (Китай), SAMRC (ЮАР) в рамках научного проекта № 20-51-80002.

Введение

Тема отнесения объектов интереса к некоторым обозначенным классам с использованием методов машинного обучения имеет обширное применение в различных сферах научных исследований и производства. При моделировании распространения инфекционных заболеваний приходится иметь дело с существенной неопределенностью, которую следует учитывать в задачах поддержки принятия решений на основе результатов вычислительных экспериментов. Однако методы кластерного анализа в основном представляют собой четкое разбиение множества элементов, при котором каждый объект строго принадлежит к определенному кластеру. На практике свойства объекта могут иметь неоднозначный характер из-за погрешностей при сборе данных, сильной зашумленности данных, а также в силу ограниченности выборки, поэтому в задаче кластеризации можно применить разделение на нечеткие множества для улучшения качества прогнозирования модели, а также для составления полной вероятностной картины принадлежности ко всем кластерам. В работе [1] рассматривается аналогичный подход в задаче кластеризации измерительных преобразователей давления. Значимым аспектом при кластеризации объектов интереса является важность кластеров для конечного пользователя. Так, к примеру, при решении задачи детекции дефектов деталей на производстве разные степени поврежденности детали (кластеры) имеют разные приоритет и степень важности.

Математическим аппаратом, позволяющим определить вероятность принадлежности объекта к кластеру и провести ранжирование с учетом важности всех кластеров, является нечеткая логика. Теория нечетких множеств была предложена Заде в 1965 году и предоставляет мощный инструмент для нечеткой кластеризации [2]. Нечеткий кластерный анализ позволяет вычислить вероятность принадлежности объекта к кластеру, что способствует учету неопределённостей при принятии решений на основе приоритетов соответствующих кластеров.

В данной работе исследуются алгоритмы нечетких вычислений в задаче кластеризации объектов интереса. Цель работы – реализовать нечеткий кластеризатор с ранжированием по важности кластеров. Обученную модель можно применять для решения широкого спектра задача в силу её универсальности и возможности обучаться на новых данных.

В разработанной системе входными данными является вектор с набором признаков, определяющих некоторые параметры объектов. Для тестирования реализованной системы в работе используются координаты некоторых объектов интереса, соответствующие позам человека в пространстве. Получение степени важности объекта происходит в два этапа: нечеткой кластеризацией вычисляется вектор вероятностей принадлежности объекта к каждому из классов, далее происходит ранжирование объекта по степени важности с использованием аппарата нечетких вычислений путем обработки наибольшей из полученных ранее вероятностей.

Постановка задачи кластеризации

Целью рассматриваемых алгоритмов машинного обучения является решение обратной задачи моделирования: поиск параметров модели, которая помогает принимать решения, предоставлять прогнозы на основе новых данных, путем поиска взаимосвязей и закономерностей в данных. Далее обученной модели передаются новые данные, на основе которых модель строит прогноз, возвращает значения функции принадлежности для всех номеров кластеров [3]. Кластеризация применяется для решения задачи группировки множества объектов на кластеры так, что элементы, принадлежащие одному кластеру, обладают схожими свойствами и отличны от элементов из других кластеров. Иначе разбиение множества наблюдений на подгруппы происходит таким образом, что объекты внутри подгруппы ближе всего к центроиду – центру тяжести этого кластера. Задача кластеризации является задачей обучения без учителя.

Кластеризация данных выполняется путем оценки близости элементов множества X. Это означает, что объекты размещены в топологическом пространстве, а близость измеряется с помощью несходства между двумя объектами. Несхожесть определяет количественно близость между двумя точками, поэтому чем меньше мера несхожести между двумя объектами, тем они ближе.

Задача кластеризации состоит в том, чтобы множество объектов X, на котором задано расстояние между элементами $\rho(x, x')$, разбить на кластеры, чтобы объекты внутри каждого подмножества были близки относительно метрики ρ, то есть сопоставить каждому элементу $x_i \in X^m$ из выборки $X^m = \{x_1, \dots, x_m\} \subset X$ метку $y_i \in Y$ из Y множества идентификаторов кластеров, где метрикой чаще всего является евклидово расстояние, определяемое следующим образом:

$$\rho(x, x') = \left|\left|x - x'\right|\right| = \sqrt{\sum_{p=1}^{n}(x_p - {x'}_p)^2}. \quad (1)$$

При этом множество меток кластеров в некоторых случаях известно заранее или же необходимо выяснить оптимальное число кластеров.

Стоит помнить, что результат разбиения на кластеры неоднозначен, необходимо оценивание результата кластеризации, на который влияют множество параметров, например количество кластеров, которое часто выбирается субъективно.

Кластеризация методом k-средних является одним из самых используемых методов разбиения множества на заданное количество k кластеров $S_1, \dots, S_k$.

Алгоритм метода k-средних:

1. Выбрать случайным образом k точек $\mu_i, i = 1, \dots, k$ из выборки в качестве начальных центров кластеров.

2. Рассчитать евклидово расстояние от каждого центроида до остальных элементов выборки. Для каждого объекта выбрать наименьшее расстояние до центра кластера.

$$x_i \in S_j \Leftrightarrow j = \arg\min_{k} \rho(x_i, \mu_k)^2. \quad (2)$$

3. Пересчитать центры кластеров:

$$\mu_k = \frac{1}{|S_i|} \sum_{x \in S_i} x. \quad (3)$$

4. Итеративно повторять пункты 2 и 3 до момента, пока центроиды кластеров не перестанут смещаться. Опционально ограничить количество итераций. В каждом новом расчете происходит изменение границ кластеров и изменение их центров.

Свойства метода k-средних:

- вычислительная мощность равна ki, где k – число кластеров, i – число итераций [4];
- алгоритм недетерминированный, неустойчивый;
- алгоритм прост в реализации;
- количество кластеров задается как параметр;
- алгоритм чувствителен к выбросам и шуму, к начальным условиям;
- алгоритм неприменим к категориальным данным.

Данный алгоритм разбивает множество наблюдений на четкие подмножества без учета того, что по некоторым свойствам элементы могут быть также отнесены и другим кластерам. Метод k-средних проводит четкую границу между подмножествами элементов, однозначно сопоставляя объекту метку из множества кластеров, что не всегда является объективным и не всегда подходит для решения задач на реальных данных.

Нечеткая кластеризация

Рассмотрим метод нечеткой кластеризации C-средних, при котором каждому элементу из выборки рассчитывается степень его принадлежности к каждому из заданных кластеров в диапазоне от 0 до 1. Таким образом происходит применение нечетких множеств и формируются пересекающиеся кластеры [5]. Этот подход является расширением метода k-средних, позволяющим получить более гибкие результаты кластеризации, но он также требует, чтобы количество кластеров было предопределено.

Задача данного метода состоит в том, чтобы разбить множество наблюдений X мощностью n на c нечетких кластеров с центрами μ так, чтобы функция потерь стремилась к минимуму:

$$\min_{c} \sum_{i=1}^{n} \sum_{j=1}^{c} {w_{ij}}^m \rho(x_i, \mu_j)^2. \quad (4)$$

При этом на полученную задачу оптимизации накладываются условия:

$$w_{ij} \in [0,1] \text{для } \forall i,j; 0 < \sum_{j=1}^{n} w_{ij} < n \text{ для } \forall i. \quad (5)$$

Число m определяет нечеткость кластеров, от которой в итоге будут зависеть значения в матрице весов W. Чем выше значение параметра m, тем более нечеткое размытие между кластерами. При $m \to \infty$ элементы матрицы нечеткого разбиения $w_{ij} = \frac{1}{c}$; иначе при $m \to 1$ значения матрицы весов W будут сходиться либо к 1, либо к 0, при этом четко ставится в соответствие к каждому элементу выборки X номер кластера, что является задачей метода k-средних. Для задач нечеткой кластеризации достаточно $m = 2$.

Алгоритм нечеткой кластеризации:

1. Инициализировать случайным образом матрицу весов
$$W = \{w_{ij}\}, i = 1,\dots,n, j = 1,\dots,c\ ; \quad (6)$$
2. Вычислить центроиды по формуле
$$\mu_j = \frac{\sum_{i=1}^{n} {w_{ij}}^m x_i}{{w_{ij}}^m}, j = 1,\dots,c\ ; \quad (7)$$
3. Рассчитать евклидово расстояние от каждого элемента x_i до всех центроидов;
4. Провести расчет матрицы принадлежности по формуле
$$w_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{||x_i - \mu_j||}{||x_i - \mu_k||})^{\frac{2}{m-1}}}, i = 1,\dots,n, j = 1,\dots,c\ ; \quad (8)$$
5. Вычислить заново центры кластеров по формуле из пункта 2;
6. Итеративно повторять пункты 3-5, пока матрица весов на шаге t не будет удовлетворять следующему условию: $||W_t - W_{t-1}|| < \varepsilon$.

Свойства метода c-средних:

- нечеткая граница разбиения на кластеры отражает более естественное поведение данных;
- параметр m при расчете центроидов усиливает влияние элементов с большим значением степени принадлежности;
- метод c-средних подвержен шуму и зависит от выбора количества кластеров, а также первоначально инициализированных весовых значений;
- на каждой итерации в этом методе используются все данные, поэтому время расчета для большого датасета может значительно увеличиться из-за того, что оно пропорционально квадрату количества кластеров;
- при неудачных начальных параметрах может сойтись к локальному экстремуму и не найти оптимального решения [6].

После получения вектора вероятностей принадлежности объекта к каждому классу в поставленной задаче надо понять важность объекта. Поэтому нам необходим математический аппарат, который позволит выполнить операцию ранжирования. Это расчет оценки величины, при котором для неё вычисляется порядковое место на заданной шкале среди

других величин последовательности. В нашем случае степень важности объекта варьируется от 0 до 1.

Пусть E – универсальное множество, X – элемент E, а R – некоторое свойство. Обычное (четкое) подмножество A универсального множества E, элементы которого удовлетворяют свойству R, определяется как множество упорядоченных пар $A = \{\mu_A(x)/x\}$, где $\mu_A(x)$– функция принадлежности, принимающая значения в некотором вполне упорядоченном множестве M (например, $M = \{0, 1\}$). Функция принадлежности указывает степень (или уровень) принадлежности элемента x подмножеству A. Множество M называют множеством принадлежностей. Если $M = \{0, 1\}$, то нечеткое подмножество A может рассматриваться как обычное или четкое множество [8]. Нечеткое множество отличается от обычного тем, что для элементов X из подмножества E нет однозначного ответа (да - нет) относительно свойства R. В связи с этим нечеткое подмножество A универсального множества E определяется как множество упорядоченных пар. Функция принадлежности $\mu_A(x)$ указывает степень (уровень) принадлежности элемента X подмножеству A.

Величина $\sup \mu_A(x)$ – верхняя граница функции принадлежности множества A – называется высотой нечеткого множества. Нечеткое множество нормально, если $\sup \mu_A(x) = 1$. При $\sup \mu_A(x) < 1$ – нечеткое множество субнормальное. Если функция принадлежности равна 0, то нечеткое множество пусто. Если $\mu_A(x) = 1$ только для одного элемента X, то нечеткое множество унимодально. Если подмножество содержит функции принадлежности строго больше 0, то оно называется A-носителем нечеткого подмножества.

Для задания функции принадлежности используют графики в виде ломаных линий. При этом на практике встречаются треугольная, трапецеидальная и гауссова формы функции принадлежности [9]. Треугольная функция принадлежности M определяется тройкой чисел (a, b, c), и ее значения в точке x вычисляются согласно выражению (9). Пример визуализации треугольной функции принадлежности представлен на рисунке 1.

$$MF(x) = \begin{cases} 1 - \dfrac{b - x}{b - a}, a \le x \le b \\ 1 - \dfrac{x - b}{c - b}, b \le x \le c \\ 0, \text{иначе} \end{cases} \quad (9)$$

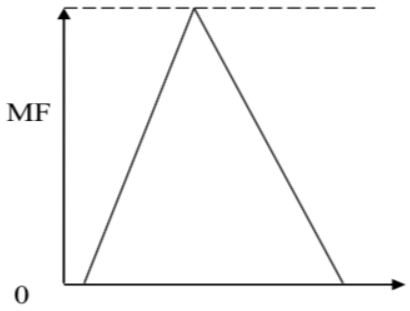

Рис. 1. Треугольная функция принадлежности

Нечеткий логический вывод по алгоритму Мамдани выполняется по нечёткой базе знаний:

$$\bigcup_{p=1}^{k_j}\left(\bigcap_{i=1}^{n} x_i = a_{i,jp} \text{ с весом } w_{jp}\right) \rightarrow y = d_j\ , \tag{10}$$

$$y = \frac{\mu_{d_1}(X^*)}{d_1} + \cdots + \frac{\mu_{d_m}(X^*)}{d_m}, \tag{11}$$

где $\mu_{jp}(x_i)$ – функция принадлежности входа нечеткому терму $a_{i,jp}$,

$\mu_{dj}(y)$ – функция принадлежности входа нечеткому терму d_j,

d_j – нечеткие термы, $X^* = (x_1^*, \dots, x_n^*)$ – входной вектор, y – нечеткое множество [10].

Наиболее часто используются следующие реализации: для операции ИЛИ – нахождение максимума и для операции И – нахождение минимума. Для перехода от нечеткого множества, заданного на универсальном множестве нечетких термов, к нечеткому множеству на интервале необходимо: "срезать" функции принадлежности; объединить (агрегировать) полученные нечеткие множества. Четкое значение выхода y, соответствующее входному вектору X^*, определяется в результате дефаззификации нечеткого множества. В данной работе была применена дефаззификация по методу центра тяжести, как наиболее распространенный подход.

Применение метода

На вход разработанной программы подается множество объектов $O = \{o_1, .., o_i\}, i = \overline{1, n}$, где каждый объект состоит из вектора признаков, описывающих основные особенности объекта $o_i = (f_1, f_2, \dots, f_m)$. При этом действует следующее ограничение – вектор признаков должен иметь одинаковую размерность для всех объектов множества O, и каждый признак f_j должен отвечать за ту же особенность объекта для каждого объекта множества. Таким образом, на вход программы подается матрица признаков объектов $I_{n\times m}$.

$$I_{n\times m} = \begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nm} \end{pmatrix}. \quad (12)$$

Датасет, выбранный для тестирования разработанной программы, находится в открытом доступе и был получен путем конвертации видео с движениями человека в массив точек с использованием библиотеки языка программирования Python – OpenPose. В нем присутствуют 4 кластера, обозначающих тип движения, который наблюдаемый объект совершает на видео. При реализации тестирования выборка была разделена на тестовую и обучающую для осуществления контроля. Модель нечеткой кластеризации была обучена на тестовых данных, полученная метрика точности (accuracy) на тестовых данных равна 34.6%, причем при использовании стандартной модели кластеризации, встроенной в библиотеку Python 3.9, точность на тех же данных составила 21.3%. Таким образом, реализованная модель нечеткой кластеризации оказалась на 62% точнее.

Ответом программы является вектор $V = (v_1, v_2, \dots, v_n)$, который определяет возможность принадлежности объекта к каждому из классов.

В программе реализован модуль нечетких вычислений для решения задачи ранжирования. После расчета вероятности принадлежности к каждому кластеру необходимо произвести ранжирование объектов по степени важности. Далее необходимо перевести важности классов объектов в нечеткие числа. Нечеткие числа задаются ломаными на плоскости. Количество нечетких чисел, которые влияют на важность объекта, совпадает с числом классов в модели, на которые идет разделение.

Для каждого кластера предусмотрено 3 параметра треугольной функции принадлежности (9) – левая и правые границы интервала и наиболее возможная важность объекта, принадлежащего данному кластеру. Количество нечетких чисел равно количеству кластеров. Их графики можно представить в виде ломанной фигуры, которая обрезается уровнем, соответствующим степени принадлежности объекта к соответствующему кластеру, уровень определяется по формуле 12:

$$MF_i(x) = \min(MF(x), v_i). \quad (12)$$

Таким образом для каждого класса формируется фигура (трапеция), соответствующая нечеткой оценке для этого кластера. Количество фигур совпадает с количеством кластеров. На рисунке 2 изображен схематично пример таких фигур. Центр тяжести ломаной, соответствующей минимуму всех функций принадлежности нечетких чисел, и есть важность объекта:

$$\mu(x) = \min\big(MF_1(x), \dots, MF_n(x)\big). \quad (13)$$

Дефаззификация ранга объекта методом центра тяжести производится по формуле:

$$\tilde{r} = \frac{\int x\mu(x)\,dx}{\int \mu(x)\,dx}. \qquad (14)$$

Если все исходные функции принадлежности были заданы в форме ломанной линии, то полученная функция принадлежности уровня предпочтений $\mu(x)$ также представляет собой ломанную линию. В этом случае интегрирование проводится по формуле (15):

$$\tilde{r} = \frac{\sum_{l=1}^{s-1}\left(a_l \frac{y_{l+1}^3 - y_l^3}{3} + b_l \frac{y_{l+1}^2 - y_l^2}{2}\right)}{\frac{1}{2}\sum_{l=1}^{s-1}\left((y_{l+1} - y_l)\left(\mu(y_{l+1}) + \mu(y_l)\right)\right)}, \text{ где} \qquad (15)$$

s – число вершин ломанной,
y_l – координата вершины ломаной,
$\mu(y_l)$ – значение функции принадлежности в вершине ломанной,
a_l и b_l – коэффициенты уравнения отрезка ломанной $\mu(y_l)=a_l y_l+b_l$.

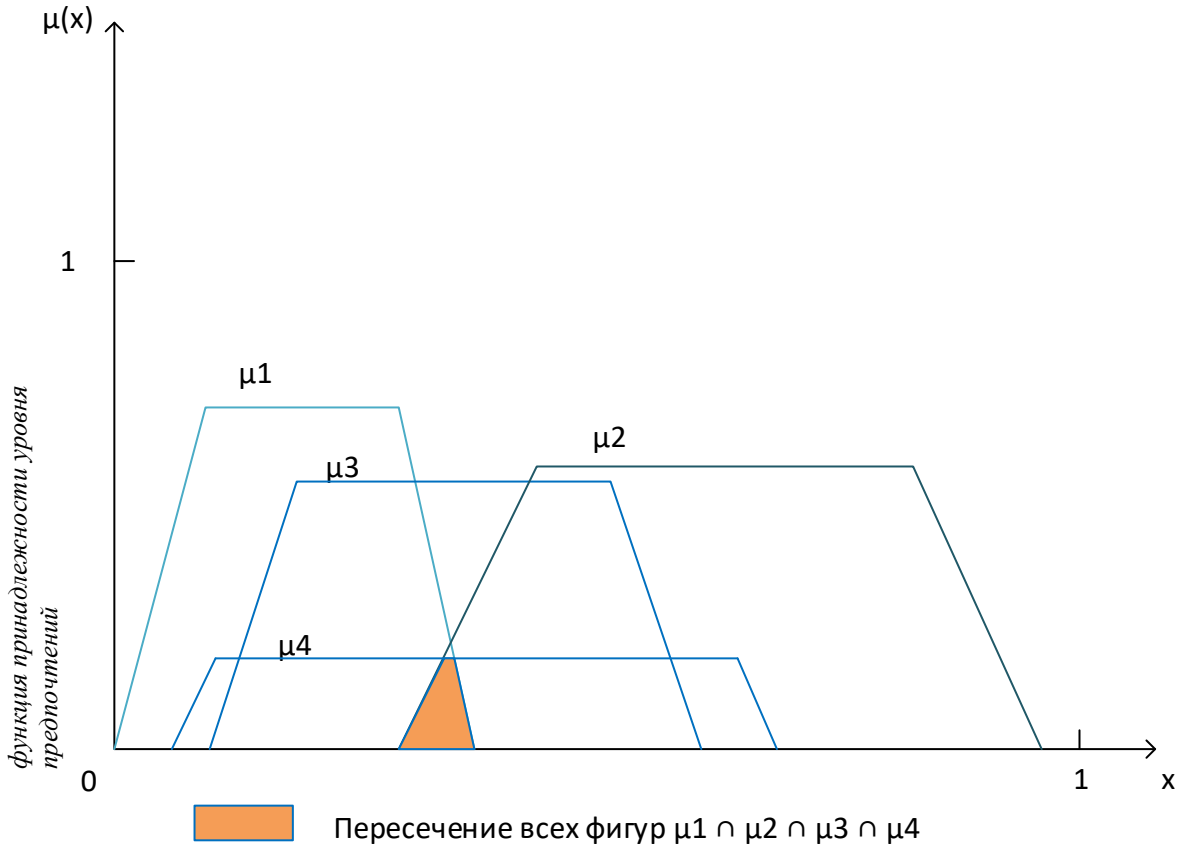

Рис. 2. Результат нечетких вычислений

После дефаззификации полученного нечеткого числа программа выводит на экран результирующий ранг объекта с учетом неопределенности и вероятности его принадлежности к каждому из кластеров:

```
In [16]: defuzzify(inferences)
         [0.0, 0.0, 0.8333333333333334, 0.16666666666666666]
Out[16]: 61.11111111111111
```

Рис. 3. Результат работы программы

Заключение

Разработана модель кластеризации объектов с последующим ранжированием важности кластеров с применением операций нечеткой алгебры. В качестве данных был использован общедоступный датасет с координатами объектов интереса.

В ходе выполнения работы было обучено несколько моделей кластеризации. Наилучшие результаты показала модель кластеризации С-средних, в рамках работы которой вычисляются вероятности принадлежности объекта ко всем кластерам модели. Такой подход помогает лучше оценить принадлежность объекта, так как пользователь может учесть приоритет объекта с оценкой вероятности и сделать вывод о соответствии прогноза действительности. В программу добавлен функционал ранжирования на основе предварительно заданной степени важности кластеров. Это позволяет агрегировать как информацию о принадлежности, так и степень важности полученного кластера для пользователя. Таким образом, внимание пользователя будет акцентировано на объекте, который представляется большую важность, с заданной вероятностью, что актуально для сферы принятия решений в области медицины.

Библиографический список

1. Лапин А.Л., Стрехнин А.И. Нечеткая кластеризация измерительных преобразователей давления // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. 2011. №23 (240).
2. Gao X., Xie W. Advances in theory and applications of fuzzy clustering // Chinese Science Bulletin. 2000. № 11.
3. Баданина Н.Д., Судаков В.А. Модели машинного обучения для классификации отзывов о банках // Препринты ИПМ им. М.В. Келдыша. 2021. № 50. 14 с. https://doi.org/10.20948/prepr-2021-50 URL: https://library.keldysh.ru/preprint.asp?id=2021-50
4. Dibya J.B., Kumar A.G. A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm // IJCTT. 2014. №10 (108-113).
5. Бротиковская Д., Зобнин Д. Алгоритм k средних (k-means). URL: https://algowiki-project.org/ru/Алгоритм_k_средних_(k-means).

6. Методы кластеризации: четкие и нечеткие. URL: https://alphacephei.com/ru/lecture7.pdf
7. Miyamoto S., Ichihashi H., Honda K. Algorithms for Fuzzy Clustering // Springer Berlin Heidelberg. 2008.
8. Songyin D. Clustering with Fuzzy C-means and Common Challenges // Journal of Physics: Conference Series, Volume 1453, 2019 2nd International Conference on Computer Information Science and Artificial Intelligence (CISAI 2019), 25-27 October 2019, Xi'an, China. 2020.
9. Определение и основные характеристики нечетких множеств. URL: http://nrsu.bstu.ru/chap21.html.
10. Обзор программного обеспечения на основе нечеткой логики. Основы теории нечетких множеств и нечеткой логики. URL: https://halzen.ru/more/obzor-programmnogo-obespecheniya-na-osnove-nechetkoi-logiki-osnovy.html.
11. Нечеткий логический вывод Мамдани. URL: https://life-prog.ru/1_12038_nechetkiy-logicheskiy-vivod-mamdani.html.

Оглавление