УДК 681.518

Л. А. Демидова, Е.И. Коняева

КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ FCM-АЛГОРИТМА НА ОСНОВЕ НЕЧЕТКИХ МНОЖЕСТВ ВТОРОГО ТИПА И ГЕНЕТИЧЕСКОГО АЛГОРИТМА

Рассматривается применение FCM-алгоритма на основе нечетких множеств второго типа для решения задачи кластеризации объектов. Для определения оптимальной комбинации значений фаззификаторов предлагается использовать генетический алгоритм.

Ключевые слова: *FCM-алгоритм, интервальные нечеткие множества второго типа, алгоритм Карника-Менделя*

Введение. Методы кластеризации на основе целевых функций используются для минимизации расстояния между образцом и прототипом кластера (таким как точка, линия, гиперэллипсоид и т.п.) и определения параметров прототипа – центра или радиуса. В дальнейшем под прототипом будем понимать точку (центр кластера). В таких подходах итерационные четкие алгоритмы типа c-средних используются для определения c разбиений, представляющих множество объектов. Если множество объектов состоит из компактных кластеров и каждый кластер разумно отделим от других, желательный результат кластеризации может быть получен с помощью алгоритма c-средних. Однако в практических задачах множества объектов редко являются такими. Достаточно часто множество объектов содержит несколько непрототипных объектов, что может привести к плохим результатам кластеризации из-за сдвига центров кластеров. Для преодоления такого нежелательного свойства четкого алгоритма c-средних применяется FCM-алгоритм (fuzzy c-means algorithm – алгоритм нечетких c-средних), использующий весовые коэффициенты (функции принадлежности) для контроля вклада объектов в определение центров кластеров. FCM-алгоритм даёт адекватные результаты кластеризации в случае, когда множество объектов содержит пересекающиеся кластеры [1, 3]. Результаты кластеризации основаны на нечетких функциях принадлежности, использующих относительные расстояния объектов относительно центров кластеров. Например, объект, расположенный далеко от центра кластера, вносит меньший вклад в процедуру поиска центров кластеров, чем объекты, расположенные близко к центру кластера.

Однако FCM-алгоритм работает хорошо, если множество объектов содержит кластеры подобного объема гиперсферической формы или подобной плотности. Если кластеры в множестве объектов имеют разную плотность (различный размер кластеров/различное количество объектов, или одинаковый размер кластеров/различное количество объектов, или различный размер кластеров/одинаковое количество объектов), то работа FCM-алгоритма может существенно зависеть от выбора фаззификатора m [3]. Если объем кластеров в множестве объектов увеличивается или число объектов в каждом кластере уменьшается, то нечеткая степень принадлежности для объектов в кластере будет изменяться. В FCM-алгоритме увеличение (уменьшение) нечеткой степени принадлежности для объектов может привести к нежелательному изменению результатов кластеризации.

При применении одного фаззификатора m предполагается, что FCM-алгоритм основан на использовании нечетких множеств первого типа (НМТ1) или просто нечетких множеств. Использование интервальных нечетких множеств второго типа (ИНМТ2) позволяет ввести в рассмотрение два фаззификатора m_1 и m_2, существенно улучшив результаты кластеризации.

1. FCM-алгоритм

FCM-алгоритм – итерационный алгоритм, вычисляющий нечеткие функции принадлежности для объектов и параметры центров кластеров в соответствии с функциями принадлежности. Функции принадлежности играют роль весовых коэффициентов, представляя степень вклада объекта в оценку центров кластеров. Размер вклада зависит от выбора фаззификатора m. При применении FCM-алгоритма определяются локально-оптимальное нечеткое разбиение, описываемое совокупностью функций принадлежности, и центры нечетких кластеров. Для получения адекватных результатов нечеткой кластеризации необходимо многократное выполнение

FCM-алгоритма при заданном числе кластеров для различных исходных нечетких разбиений для принятия окончательного решения об искомой нечеткой кластеризации [1].

FCM-алгоритм на основе НМТ1 предполагает минимизацию целевой функции:

$$J(U,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_j(x_i)\right)^m\cdot d_{ji}^2 \qquad (1)$$

при

$$\sum_{j=1}^{c}u_j(x_i)=1\ (i=\overline{1,n}), \qquad (2)$$

$$d_{ji}=\left\|x_i-v_j\right\|, \qquad (3)$$

где $U=[u_j(x_i)]$ – нечеткое c-разбиение множества объектов $\{x_i\}$ на основе функций принадлежности $u_j(x_i)$; $V=(v_1,...,v_c)$ – центры кластеров; d_{ji} – расстояние между центром кластера v_j и объектом x_i; m – фаззификатор ($m\in R$, $m>1$); c – количество кластеров; n – количество объектов; $i=\overline{1,n}$; $j=\overline{1,c}$.

FCM-алгоритм предполагает выполнение следующих шагов.

1. Инициализация начального нечеткого разбиения $U=[u_j(x_i)]$, удовлетворяющего условию (2).

2. Вычисление центров кластеров:

$$v_j=\frac{\sum_{i=1}^{n}u_j(x_i)^m\cdot x_i}{\sum_{i=1}^{n}u_j(x_i)^m}. \qquad (4)$$

3. Вычисление новых функций принадлежности:

$$u_j(x_i)=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ji}}{d_{ki}}\right)^{\frac{2}{m-1}}}. \qquad (5)$$

4. Шаги 2 и 3 повторяются до тех пор, пока не будет выполнено заданное число итераций s или не будет достигнута заданная точность $\left|J(U,V)-J'(U,V)\right|\le\varepsilon$, где $J(U,V)$, $J'(U,V)$ – значения целевой функции на двух последовательных итерациях.

2. Неопределенность фаззификатора

FCM-алгоритм дает хорошие результаты кластеризации, если кластеры идентичны по структуре и плотности. Максимально нечеткие степени принадлежности (средняя вертикальная линия на рисунках 1 и 2) там, где объекты расположены на одинаковом расстоянии от центров кластеров (относительное расстояние между объектом и каждым центром кластера равно 0,5). Эти объекты будут оказывать меньшее влияние на процедуру поиска центров кластеров, чем те, которые удалены от этой вертикальной линии. Вертикальная линия определяет границу решения: объекты, расположенные слева (справа) от нее, принадлежат кластеру A_1 (A_2).

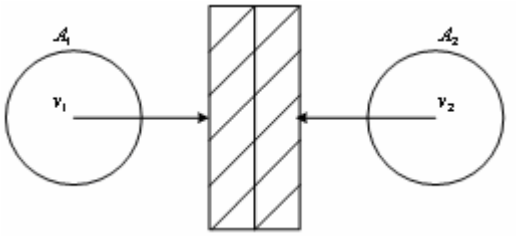

Рисунок 1 - Максимальная нечеткая область для кластеров одинакового объема с малым значением фаззификатора m

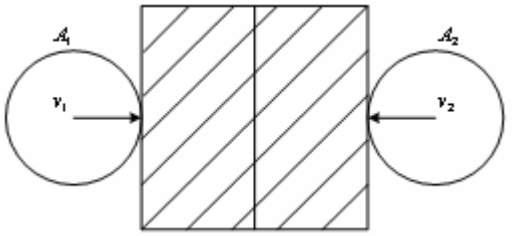

Рисунок 2 - Максимальная нечеткая область для кластеров одинакового объема с большим значением фаззификатора m

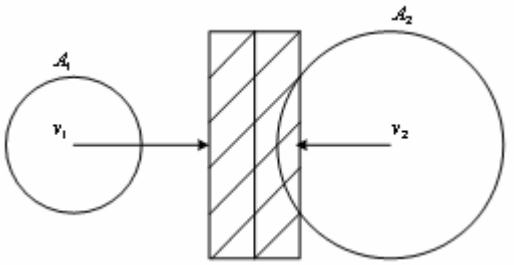

Рисунок 3 - Максимальная нечеткая область для кластеров разного объема с малым значением фаззификатора m

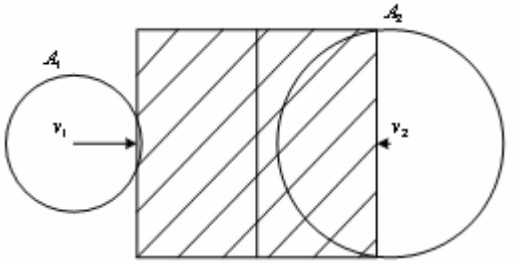

Рисунок 4 - Максимальная нечеткая область для кластеров разного объема с большим значением фаззификатора m

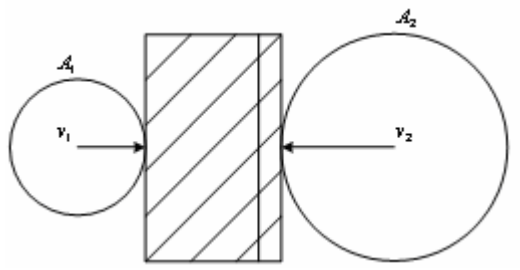

Рисунок 5 - Желательная максимальная нечеткая область для кластеров разного объема

Фаззификатор m определяет максимальную нечеткую область. Обычно хорошие результаты кластеризации получаются при $m = 2$. Для случаев, изображенных на рисунках 3 и 4, FCM-алгоритм может дать плохие результаты кластеризации при выборе несоответствующего фаззификатора m из-за различия в объеме (плотности) между двумя кластерами. Объекты, лежащие слева от максимальной нечеткой области и принадлежащие кластеру A_2, будут вносить больший вклад в кластер A_1, чем в кластер A_2. Следовательно, необходим контроль при выборе фаззификатора m для улучшения результатов кластеризации в случае, когда кластеры существенно различны в размере. При выборе большого фаззификатора m ширина максимальной нечеткой области будет, как на рисунке 4. Это может показаться желательным с точки зрения кластера A_1, но найденный центр кластера v_1 сдвинется к кластеру A_2 и отклонится от идеального центра кластера A_1. Это плохо и с точки зрения кластера A_2, так как нежелательный центр кластера A_1 может оказать влияние на кластер A_2. Идеальная ситуация – получение максимальной нечеткой области с широкой левой и узкой правой подобластями (рисунок 5). При этом для каждой правой и левой нечеткой подобластей относительное расстояние по отношению к вертикальной линии равно 0,5. Но такая максимальная нечеткая область не может быть получена с помощью FCM-алгоритма на основе НМТ1 ввиду единственности фаззификатора m. Чтобы управлять неопределенностью, существующей при задании максимальной нечеткой области в FCM-алгоритме, необходимо использовать два фаззификатора m_1 и m_2 для расширения множества объектов на ИНМТ2 [3]. Эта максимальная нечеткая область является неопределенной, так как построена с помощью двух фаззификаторов m_1 и m_2, представляющих различные степени нечеткости [3]. Поэтому необходимо рассматривать функции принадлежности для объекта как неопределенные (нечеткие) в отличие от определенных (четких) в FCM-алгоритме на основе НМТ1. Под неопределенностью нечеткой функции принадлежности объекта будем понимать неопределенную максимальную нечеткую область.

3. Использование интервальных нечетких множеств второго типа

Первичная функция принадлежности J_{x_i} объекта x_i при описании неопределенности с помощью ИНМТ2 может быть определена как интервальная функция принадлежности со всеми вторичными степенями первичных функций принадлежности, равными 1. Тогда ИНМТ2 $\widetilde{X}$ может быть представлено как [3]:

$$\widetilde{X} = \left\{\left((x,u),\mu_{\widetilde{X}}(x,u)\right)\right| \\ \forall x \in X, \forall u \in J_x \subseteq [0,1], \mu_{\widetilde{X}}(x,u) = 1\Big\}. \quad (6)$$

При определении интервальных первичных функций принадлежности объекта x_i рассмотрим нижнюю и верхнюю интервальные функции принадлежности, используя два различных значения фаззификатора m:

$$\overline{u}_j(x_i) = \begin{cases} u_j^1(x_i), & \text{если } u_j^1(x_i) > u_j^2(x_i) \\ u_j^2(x_i), & \text{если } u_j^1(x_i) \le u_j^2(x_i) \end{cases}, \quad (7)$$

$$\underline{u}_j(x_i) = \begin{cases} u_j^1(x_i), & \text{если } u_j^1(x_i) \le u_j^2(x_i) \\ u_j^2(x_i), & \text{если } u_j^1(x_i) > u_j^2(x_i) \end{cases}, \quad (8)$$

где $u_j^p(x_i) = \dfrac{1}{\sum\limits_{k=1}^{c}\left(\dfrac{d_{ji}}{d_{ki}}\right)^{\frac{2}{m_p-1}}}$, $p = 1,2$.

Использование фаззификаторов m_1 и m_2, представляющих различные степени нечеткости, приводит к различным целевым функциям, которые должны быть минимизированы с помощью FCM-алгоритма при $m = m_1$ и $m = m_2$:

$$J_{m_1}(U,V) = \sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_j(x_i)\right)^{m_1} \cdot d_{ji}^2, \quad (9)$$

$$J_{m_2}(U,V) = \sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_j(x_i)\right)^{m_2} \cdot d_{ji}^2. \quad (10)$$

Пусть имеются 2 кластера в одномерном пространстве с центрами v_1 и v_2. На рисунке 6 приведены графики функций принадлежности, соответствующие центру v_1, для объектов, лежащих между центрами кластеров, для различных значений фаззификатора m. Степени принадлежности являются максимально четкими при $m \to 1$: объекты, размещенные слева (справа) от максимальной нечеткой области, полно-

стью принадлежат (или не принадлежат) кластеру A_1 (A_2). Степени принадлежности являются максимально нечеткими при $m \to \infty$: объекты, размещенные в центрах кластеров, полностью принадлежат (или не принадлежат) кластеру A_1 (A_2), иначе степень принадлежности равна 0,5.

На рисунке 7 приведен пример ИНМТ2 в случае двух кластеров для комбинаций фаззификаторов $m_1 = 2$ и $m_2 = 5$. Различие в функциях принадлежности определяет для ИНМТ2 «отпечаток неопределенности» FOU, закрашенный на рисунке в черный цвет [3, 5]. Степени принадлежности для объектов, размещенных в каждом текущем центре кластера (для текущей итерации), равны 1 и равны 0 для другого центра. Следовательно, для таких объектов FOU не существует. Для объектов, расположенных на одинаковом расстоянии от центров обоих (двух) кластеров, степени принадлежности равны 0,5. Для таких объектов FOU также не существует. Таким образом, не существует неопределенности для объектов с максимальной и минимальной степенями принадлежности.

Управление неопределенностью фаззификатора m, позволяющее существенно улучшить результаты кластеризации, осуществляется с помощью: вычисления центров (центроидов) кластеров и дефаззификации (получения четкого разбиения) для конечного решения о результатах кластеризации [3].

При вычислении центров кластеров с помощью FCM-алгоритма на основе ИНМТ2 используются операция «type-reduction» («понижение типа») и методы дефаззификации ИНМТ2 [3].

Центроид НМТ1 для n объектов может быть вычислен по формуле:

$$v_X = \frac{\sum_{i=1}^{n} x_i \cdot u(x_i)}{\sum_{i=1}^{n} u(x_i)}. \qquad (11)$$

По принципу расширения центроид НМТ2 $\tilde{X}$ вычисляется как [5]:

$$v_{\tilde{X}} = \sum_{u(x_1)\in J_{x_1}} \dots \sum_{u(x_n)\in J_{x_n}} F \Big/ \frac{\sum_{i=1}^{n} x_i \cdot u(x_i)}{\sum_{i=1}^{n} u(x_i)}, \qquad (12)$$

где $F = f(u(x_1)) * \dots * f(u(x_n))$.

Использование ИНМТ2 обеспечит улучшение результатов кластеризации при уменьшении вычислительной сложности по сравнению с обобщенными нечеткими множествами второго типа [5]. Для ИНМТ2 в формуле (8) можно заменить все $f(u(x_i))$ ($i = \overline{1,n}$) на 1 и использовать эту формулу для нечеткой кластеризации, введя дополнительно в формулу нечеткую степень m:

$$v_{\tilde{X}} = \sum_{u(x_1)\in J_{x_1}} \dots \sum_{u(x_n)\in J_{x_n}} 1 \Big/ \frac{\sum_{i=1}^{n} x_i \cdot u(x_i)^m}{\sum_{i=1}^{n} u(x_i)^m}. \qquad (13)$$

В результате оцениваемые центры кластеров представляются интервалом $v_{\tilde{X}} = [v_L, v_R]$ [3].

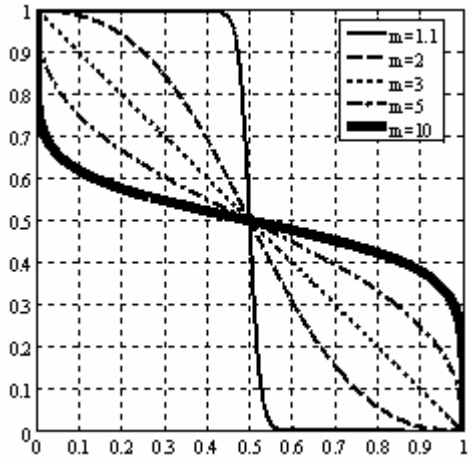

Рисунок 6 – Графики нечетких функций принадлежности для FCM-алгоритма

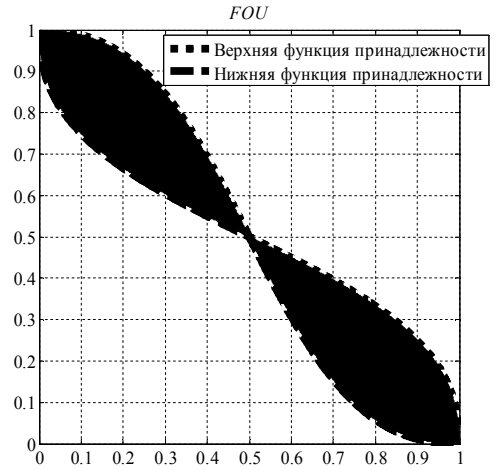

Рисунок 7 – «Отпечаток неопределенности» для комбинации $m_1 = 2$ и $m_2 = 5$

Для оценки центров кластеров v_L и v_R обычно рассматривают все «вложенные» НМТ1 для ИНМТ2, которые описываются верхними и нижними значениями первичных функций принадлежности J_{x_i}. В общем случае существует 2^n вложенных НМТ1 в ИНМТ2, описывающее n объектов. Для уменьшения вычислительной сложности при оценке центров кластеров целесообразно использовать итерационный алгоритм Карника-Менделя [3, 5].

4. Итерационный алгоритм Карника-Менделя

Итерационный алгоритм Карника-Менделя позволяет определить два «вложенных» НМТ1 – L и R – внутри FOU ИНМТ2 $\widetilde{X}$. Множества L и R имеют минимально и максимально возможные центроиды v_L и v_R в $\widetilde{X}$ соответственно. Четкое значение центроида определяется как среднее значение от центроидов НМТ1: L и R.

Пусть для каждого из объектов кластеризации x_i $(i=\overline{1,n})$ заданы значения всех q характеристик, измеренные в некоторой количественной шкале: $x_i=(x_{i1},...,x_{iq})$.

Алгоритм Карника-Менделя для поиска максимума v_R центра кластера v_j имеет вид:

1. Для множества объектов $x_i=(x_{i1},...,x_{iq})$ $(i=\overline{1,n})$ вычисляются функции принадлежности в соответствии с формулами (7) и (8).

2. Выбирается значение фаззификатора m (любое из m_1 и m_2).

3. Вычисляется центроид $v'_j=(v'_{j1},...,v'_{jq})$ по формуле (11) и $u_j(x_i)=\left(\overline{u}_j(x_i)+\underline{u}_j(x_i)\right)/2$.

4. Выполняется сортировка индексов n объектов $(i=\overline{1,n})$ по каждой характеристике l $(l=\overline{1,q})$ по возрастанию:

$x_{11},...,x_{n1}$,

……

$x_{1q},...,x_{nq}$.

5. По каждой характеристике l $(l=\overline{1,q})$ выполняется поиск индекса k $(1\le k\le n-1)$ такого, что $x_{kl}\le v'_{jl}\le x_{(k+1)l}$.

6. Для всех n объектов вычисляются функция принадлежности: если $i\le k$, то $u_j(x_i)=\underline{u}_j(x_i)$; иначе – $u_j(x_i)=\overline{u}_j(x_i)$.

7. Вычисляется центроид для v''_{jl} по формуле (4).

8. Если $v'_{jl}=v''_{jl}$, то l-я координата центра j-го кластера считается вычисленной и осуществляется переход к шагу 9. В противном случае полагается, что $v'_{jl}=v''_{jl}$, и осуществляется переход к шагу 5 для уточнения l-й координаты j-го кластера.

9. Если $l<q$, то номер координаты увеличивается на единицу: $l=l+1$ и осуществляется переход к шагу 5 для уточнения l-й координаты j-го кластера. Если $l=q$, считается, что процедура вычисления максимума v_R центра кластера v_j завершена и $v_R=v'_j=(v'_{j1},...,v'_{jq})$.

Минимум v_L центра кластера v_j вычисляется аналогичным образом с заменой действий шага 6 на следующее: если $i\le k$, то $u_j(x_i)=\overline{u}_j(x_i)$; иначе – $u_j(x_i)=\underline{u}_j(x_i)$.

Результирующее интервальное НМТ1 может быть записано в виде: $v_j=1{,}0/[v_L,v_R]$. Четкое значение для центра кластера v_j находится с помощью операции «понижения типа» [3, 5]:

$$v_j=(v_L+v_R)/2. \qquad (14)$$

«Четкое разбиение» в FCM-алгоритме на основе НМТ1 находится в соответствии с правилом:

«Если $(u_j(x_i)>u_t(x_i))$,

для $t=\overline{1,...,c}$ и $j\ne t$, (15)

то x_i относится к кластеру j».

Перед выполнением «четкого разбиения» в FCM-алгоритме на основе ИНМТ2 необходимо предварительно выполнить «понижение типа» для функций принадлежности объектов $u_j(x_i)$, используя левые и правые функции принадлежности ($u_j^L(x_i)$ и $u_j^R(x_i)$), найденные при вычислении v_L и v_R соответственно. «Понижение типа» может быть выполнено как:

$$u_j(x_i)=\frac{u_j^R(x_i)+u_j^L(x_i)}{2},\ j=\overline{1,...,c}, \qquad (16)$$

где $u_j^R(x_i)=\dfrac{\sum_{l=1}^{q}u_{jl}(x_i)}{q}$ при

$$u_{jl}(x_i)=\begin{cases}\overline{u}_j(x_i), \text{если } x_{il} \text{ использует } \overline{u}_j(x_i) \text{ для } v_j^R\\ \underline{u}_j(x_i), \text{иначе}\end{cases};$$

$u_j^L(x_i)=\dfrac{\sum_{l=1}^{q}u_{jl}(x_i)}{q}$ при

$$u_{jl}(x_i)=\begin{cases}\overline{u}_j(x_i), \text{если } x_{il} \text{ использует } \overline{u}_j(x_i) \text{ для } v_j^L\\ \underline{u}_j(x_i), \text{иначе}\end{cases}.$$

Тогда «четкое разбиение» выполняется в соответствии с правилом (15).

Вычислительная сложность FCM-алгоритма на основе ИНМТ2 не превышает $O(n^3)$, а для ряда практических случаев может быть оценена как $O(n^2)$ [3].

FCM-алгоритм на основе ИНМТ2 является обобщением FCM-алгоритма на основе НМТ1 при $m_1=m_2$.

5. Генетический алгоритм поиска оптимальной комбинации значений фаззификаторов

Для поиска оптимальной комбинации значений фаззификаторов m_1 и m_2 следует использовать генетический алгоритм [2], так как перебор всевозможных наиболее часто используемых значений фаззификаторов m_1 и m_2 не всегда не приносит желаемый результат.

Для решения задачи поиска оптимальной комбинации значений фаззификаторов m_1 и m_2 хромосома может быть задана в виде: $s=(m_1,m_2)$, где $m_1,m_2 \in (1,\ m_{max})$; m_{max} – некоторое действительное число, определяющее максимальное значение фаззификатора; $m_1 < m_2$.

В большинстве практических задач m_{max} удовлетворяет неравенству: $m_{max} \le 30$.

Анализ качества кластеризации на основе различных индексов кластеризации (таких как коэффициент разбиения PC, энтропия разбиения PE, индекс Хие-Бени XB, индекс плотности CS [4]) на различных тестовых примерах показал, что наилучшие результаты кластеризации обеспечивает индекс плотности CS, который должен быть максимизирован:

$$CS=\frac{s_1}{s_2}, \tag{17}$$

$$s_1=\sum_{j=1}^{c}\left(\frac{1}{|X_j|}\cdot\sum_{x_i\in X_j}\max_{x_r\in X_j}\left(d(x_i,x_r)\right)\right), \tag{18}$$

$$s_2=\sum_{j=1}^{c}\left(\min_{\substack{t=1,c\\ t\neq j}}\left(d(v_j,v_t)\right)\right), \tag{19}$$

$$d(x_i,x_r)=\|x_i-x_r\|, \tag{20}$$

$$d(v_j,v_t)=\|v_j-v_t\|. \tag{21}$$

В связи с этим индекс плотности CS был выбран в качестве функции соответствия для генетического алгоритма.

Применение в качестве целевых функций индексов, использующих в своей записи значение фаззификатора m, оказалось невозможно, во-первых, ввиду наличия двух фаззификаторов m_1 и m_2, а, во-вторых, из-за того, что само значение фаззификатора существенным образом влияет на значение индекса кластеризации (при поиске экстремума).

Выбор родителя заключается в выборе хромосомы, максимизирующей индекс плотности CS по формуле (17), из двух случайно выбранных. Затем выбранные таким образом хромосомы-родители используются в операции скрещивания.

При выполнении операции скрещивания выбирается вероятность скрещивания P_c и генерируется случайное число N_c. Если $P_c > N_c$, то случайным образом выбирается точка скрещивания z и выполняется скрещивание.

При выполнении операции мутации выбирается вероятность мутации P_m и генерируется случайное число N_m. Если $P_m > N_m$, то случайным образом выбирается точка мутации z и выполняется мутация.

Тогда генетический алгоритм имеет вид.

1. Случайным образом создается популяция размера P. При этом выполняется проверка условия: $m_1 < m_2$.

2. При $g < G$ (G и g – максимальное и текущее количество генераций генетического алгоритма соответственно) вычисляется функция соответствия (17) для каждой хромосомы и создается $P/2$ пар хромосом-родителей.

3. Выполняются операции скрещивания и мутации для текущей популяции. При этом выполняется проверка условия: $m_1 < m_2$.

4. Создается новая популяция размера P, дополненная хромосомами-детьми в количестве $P * P_c$, затем $P * P_c$ хромосом с худшими значениями функции соответствия (17) отбрасываются. При $g < G$ осуществляется переход к шагу 2.

5. Выбирается лучшая хромосома, которая максимизирует функцию соответствия (17). Для каждого объекта определяется его принадлежность к нечетким кластерам.

6. Экспериментальные результаты

Ниже приведены два примера и сравнительные результаты кластеризации, полученные на основе НМТ1 и ИНМТ2.

В целом FCM-алгоритм работает хорошо для множеств объектов, содержащих разбиения подобного объема и подобного количества объектов.

Изменяя объем кластеров в множестве объектов, оценим эффективность предлагаемого подхода по сравнению с классическим FCM- алгоритмом на основе НМТ1.

Первый пример (рисунки 8, 9 и 10) иллюстрирует влияние выбора фаззификатора m на результаты кластеризации множества объектов, содержащего два кластера разного объема с одинаковым количеством объектов (по 9 объектов в кластере). Объекты первого, второго кластеров и центры кластеров помечены круглыми, квадрат-

ными и треугольными маркерами соответственно.

FCM-алгоритм на основе НМТ1 при $m = 1{,}1$ дает ошибку кластеризации в 3 объекта, при $m = \overline{2,...,8}$ – в 2 объекта, а при $m = \overline{9,...,30}$ – в 1 объект. Ошибочно относятся к первому (нижнему) кластеру один, два или три объекта второго (верхнего) кластера, расположенные к нему ближе, чем другие объекты второго кластера.

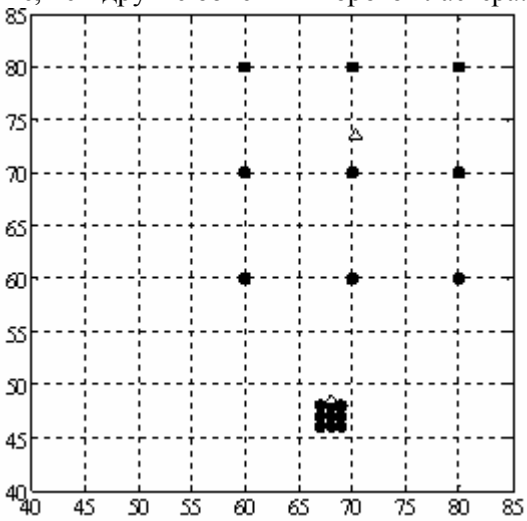

Рисунок 8 – Расположение центров кластеров для множества объектов разного объема при $m = 2$

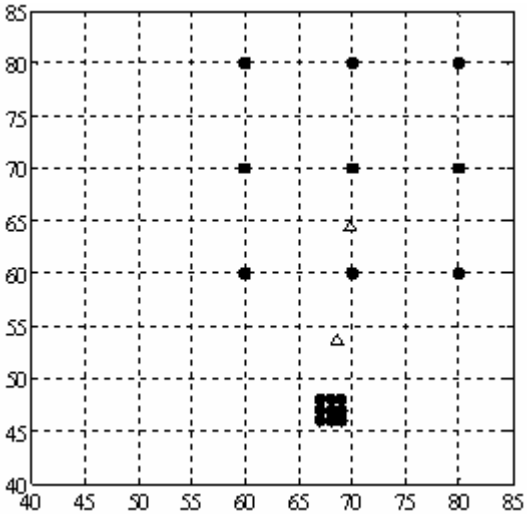

Рисунок 9 – Расположение центров кластеров для множества объектов разного объема при $m_1 = 5$ и $m_2 = 22$

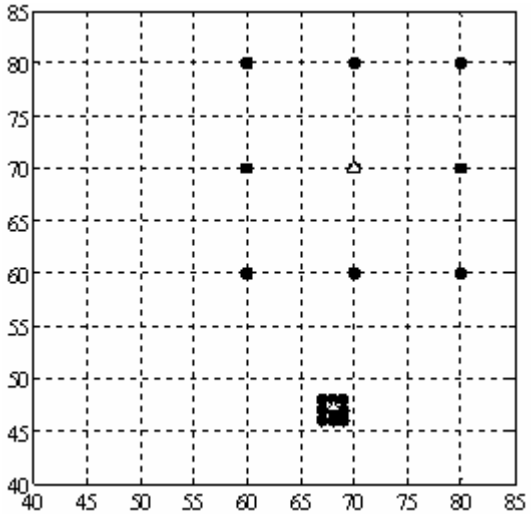

Рисунок 10 – Расположение центров кластеров для множества объектов разного объема при $m_1 = 24{,}39791$ и $m_2 = 25{,}915252$

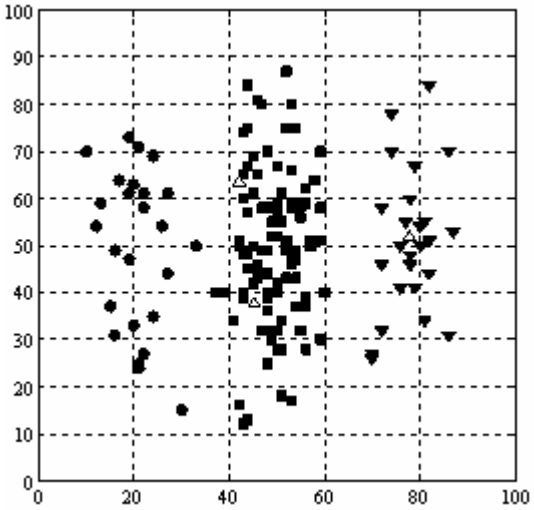

Рисунок 11 – Расположение центров кластеров для множества объектов, содержащего кластеры с различным количеством элементов при $m = 2$

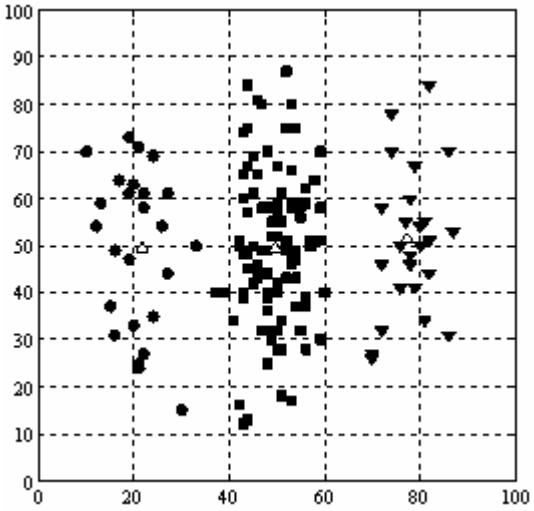

Рисунок 12 – Расположение центров кластеров для множества объектов, содержащего кластеры с различным количеством элементов

при $m_1 = 1{,}115285$ и $m_2 = 1{,}276459$

Рисунок 8 показывает расположение центров кластеров для наилучшего возможного результата кластеризации с помощью FCM-алгоритма на основе НМТ1 при фаззификаторе $m = 2$, который обычно используется для FCM-алгоритма, с ошибкой кластеризации в 2 объекта. FCM-алгоритм на основе ИНМТ2 позволяет уменьшить ошибку кластеризации, а для ряда комбинаций значений фаззификаторов m_1 и m_2 обеспечивает нулевую ошибку кластеризации. Так, для комбинаций значений фаззификаторов $m_1 = 5$ и $m_2 = 22$ кластеризация выполняется с нулевой ошибкой. Но расположение вычисленных центров следует признать неудачным (рисунок 9). Применение генетического алгоритма совместно с FCM-алгоритмом на основе ИНМТ2 позволило определить оптимальную комбинацию значений фаззификаторов $m_1 = 24{,}39791$ и $m_2 = 25{,}915252$, обеспечивающую не только нулевую ошибку кластеризации, но и ожидаемое расположение центров кластеров (рисунок 10).

Второй пример демонстрирует возможности совместного применения FCM-алгоритма на основе ИНМТ2 и генетического алгоритма для случая множества объектов, содержащего три кластера с разным количеством объектов (рисунки 11 и 12). Объекты первого, второго и третьего кластеров помечены круглыми, квадратными и треугольными «чёрными» маркерами, центры кластеров – треугольными «белыми» маркерами. При этом объекты первого и третьего кластеров содержат по 25 элементов, второй кластер содержит 100 элементов, а характеристики объектов в каждом кластере распределены в соответствии с нормальным законом распределения $N(c, \sigma^2)$, где c – математическое ожидание, σ – среднеквадратическое отклонение.

Формула плотности вероятности случайной величины x, распределенной по нормальному закону, имеет вид:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{(x-c)^2}{2 \cdot \sigma^2}}. \qquad (22)$$

Нормальное распределение с заданными c и σ может быть получено из нормализованного нормального распределения $N(0,1)$ как:

$$z = c + \sigma \cdot x, \qquad (23)$$

где x – случайная величина с нормализованным нормальным законом распределения.

Предположим, что данные в первом кластере распределены по первой и второй характеристикам по законам $N(20,36)$ и $N(50,256)$ соответственно; во втором кластере – по законам $N(50,36)$ и $N(50,256)$ соответственно; в третьем кластере – по законам $N(80,36)$ и $N(50,256)$ соответственно. Так как по каждой характеристике данные распределены по нормальному закону $N(c, \sigma^2)$, то очевидно, что центры вычисленных кластеров должны быть близки к точкам с координатами $(20,50)$, $(50,50)$, $(80,50)$ для первого, второго и третьего кластеров соответственно.

Результаты кластеризации с помощью FCM-алгоритма на основе НМТ1 при различных значениях фаззификатора m оказываются далеки от ожидаемых. Это можно объяснить тем, что кластеры содержат разное количество объектов. Рисунок 11 иллюстрирует расположение центров кластеров для фаззификатора $m = 2$. Совместное применение генетического алгоритма и FCM-алгоритма на основе ИНМТ2 позволило определить комбинацию значений фаззификаторов m_1 и m_2, обеспечивающую максимальную близость к предполагаемым центрам кластеров. Также найденная комбинация значений фаззификаторов $m_1 = 1{,}115285$ и $m_2 = 1{,}276459$ обеспечивает нулевую ошибку кластеризации, то есть все объекты оказываются отнесенными к «своим» кластерам (рисунок 12).

Приведенные выше примеры демонстрируют преимущества совместного применения генетического алгоритма и FCM-алгоритма на основе ИНМТ2 при поиске оптимальной комбинации значений фаззификаторов m_1 и m_2, обеспечивающей наилучшие результаты кластеризации.

Выводы

Использование ИНМТ2 и введение комбинации значений фаззификаторов m_1 и m_2 позволяют представлять и управлять неопределенностью, которая возникает при анализе множества объектов, образованного из кластеров различной плотности (различного объема с различным числом объектов). В результате удается определить более точное положение центров кластеров и, следовательно, улучшить результаты нечеткой кластеризации.

Применение генетического алгоритма позволяет найти оптимальную комбинацию значений фаззификаторов m_1 и m_2, обеспечивающую лучшие результаты нечеткой кластеризации, что подтверждается максимальным значением индекса плотности CS для данного множества объектов при заданном числе кластеров c.

Библиографический список

1. *Демидова Л.А.*, *Кираковский В.В.*, *Пылькин А.Н.* Алгоритмы и системы нечеткого вывода в задачах диагностики городских инженерных коммуникаций. – М.: Радио и связь, Горячая линия – Телеком, 2005. 592 с., ил.

2. *Ярушкина Н.Г.* Основы теории нечетких и гибридных систем: учеб. пособие. – М.: Финансы и статистика, 2004. 320 с.: ил.

3. *Hwang C.*, *Rhee F.C.-H.* uncertainfuzzy clustering: interval type-2 fuzzy approach to c-means // IEEE Trans. on Fuzzy Systems. 2007. vol. 15. № 1. – P. 107-120.

4. *Galda H.* Development of segmentation method for dermoscopic images based on color clustering. Kobe University, Graduate school of science and technology. 2003. – 79 p.

5. *Mendel Jerry M.* Type-2 fuzzy sets and systems: an overview. // IEEE Computational intellegence magazine. 2007. vol. 2. № 1 – P. 20-29.