УДК 519.254

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

Демидова Л. А. – канд. техн. наук, доц. кафедры «Вычислительная и прикладная математика», тел.: (4912) 46-03-64, e-mail: liliya.demidova@rambler.ru; *Кираковский В. В.* – канд. техн. наук, доц. кафедры «Вычислительная и прикладная математика», тел.: (4912) 46-03-64, e-mail: mail@pgproject.ru; *Коротаев А. Н.* – соиск., электроник кафедры «Вычислительная и прикладная математика», тел.: (4912) 46-03-64, e-mail: graph2@rambler.ru (РГРТУ)

Для решения задач статистической обработки и обучения без учителя часто используют кластерный анализ. Среди методов кластеризации наиболее популярным является метод *c*-средних. Использование аппарата теории нечетких множеств совместно с методом *c*-средних, а именно, алгоритма нечетких c-средних, дает хорошие результаты. В данной статье рассматривается решение задачи кластеризации объектов с использованием алгоритма возможностных *c*-средних совместно с интервальными нечёткими множествами второго типа и генетическим алгоритмом. Показано, что этот метод демонстрирует хорошие результаты, что говорит о целесообразности его использования при решении задач кластеризации.

For problems of statistic processing and teaching without teacher cluster analysis is often used. Among clustering methods the most popular is c-means method. Using fuzzy-set theory together with *c*-means method, namely, fuzzy *c*-means algorithm, gives good results. In this article the problem of clustering objects with using possible *c*-means together with interval type 2 fuzzy set and genetic algorithm is considered. It is shown that this method provides good results. Therefore it is expedient to make use of it for clustering problems.

Ключевые слова: кластеризация, нечеткие множества, генетический алгоритм.

Введение

Методы кластеризации на основе целевых функций используются для минимизации расстояния между образцом и прототипом кластера и определения параметров прототипа – центра или радиуса кластера. В дальнейшем под прототипом будем понимать точку (центр кластера).

Если множество объектов состоит из компактных кластеров и каждый кластер разумно отделим от других, искомый результат кластеризации может

быть получен с помощью алгоритма c-средних. Однако в практических задачах множества объектов редко являются такими. Наиболее известными алгоритмами кластеризации, основанными на учете того или иного вида неопределенности, являются алгоритм нечетких c-средних (fuzzy c-means – FCM-алгоритм) и алгоритм возможностных c-средних (possibilistic c-means – PCM-алгоритм).

FCM-алгоритм работает хорошо, если множество объектов содержит кластеры подобного объема гиперсферической формы или подобной плотности. Если объем кластеров увеличивается или количество 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. PCM-алгоритм позволяет улучшить результаты кластеризации, полученные с помощью FCM-алгоритма, в случае, если множество объектов содержит атипичные объекты (объекты-шумы), за счет ослабления свойства кластерной относительности и учета свойства типичности. Однако результаты кластеризации с использованием этого алгоритма также сильно зависят от выбора фаззификатора m и значений «ширины зоны» η_j ($j = \overline{1,c}$), если кластеры имеют существенно разную плотность или существенно разный объем.

Таким образом, можно говорить о существовании неопределенности при определении типичности объекта кластеру в PCM-алгоритме и о необходимости в реализации метода расчета значений степеней типичности объектов кластерам за счет управления неопределенностью в выборе параметров PCM-алгоритма – значений фаззификатора m значений или «ширины зоны» η_j ($j = \overline{1,c}$).

Расширение множества объектов на интервальные нечеткие множества второго типа для PCM-алгоритма

Неопределенность в PCM-алгоритме на основе интервальных нечетких множеств второго типа (ИНМТ2) может быть выражена как с помощью двух различных значений фаззификатора m (как в FCM-алгоритме на основе ИНМТ2) при фиксированном значении «ширины зоны» η_j ($j = \overline{1,c}$) для каждого кластера, так и с помощью двух различных значений «ширины зоны» η_j ($j = \overline{1,c}$) для каждого кластера при фиксированном значении фаззификатора m [1].

Пусть неопределенность в PCM-алгоритме на основе ИНМТ2 задана с помощью двух различных значений фаззификатора m: m_1 и m_2 (как в FCM-алгоритме на основе ИНМТ2) при фиксированном значении «ширины зоны» η_j ($j = \overline{1,c}$) для каждого кластера.

Если неопределенность в РСМ-алгоритме на основе ИНМТ2 выражена с помощью двух различных значений «ширины зоны» η_j ($j=\overline{1,c}$) для каждого кластера при фиксированном значении фаззификатора m, то при определении интервальных первичных функций типичности объекта x_i в РСМ-алгоритме на основе ИНМТ2 «нижняя» и «верхняя» интервальные функции типичности (для двух различных значений «ширины зоны» η_j: η_{j1} и η_{j2}) могут быть представлены как [2]:

$$\overline{w}_j(x_i)=\begin{cases}\dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}}, \text{если } \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}} > \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}} \\ \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}}, \text{если } \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}} \le \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}}\end{cases}, \quad (1)$$

$$\underline{w}_j(x_i)=\begin{cases}\dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}}, \text{если } \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}} \le \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}} \\ \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}}, \text{если } \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j1}}\right)^{\frac{2}{m-1}}} > \dfrac{1}{1+\left(\dfrac{d_{ji}}{\eta_{j2}}\right)^{\frac{2}{m-1}}}\end{cases}, \quad (2)$$

где η_{j1} и η_{j2} – две «ширины зоны», определяющие различные расстояния, на которых возможностная степень типичности равна 0,5.

Использование различных значений для фаззификатора m определяет различные целевые функции в РСМ-алгоритме на основе ИНМТ2 при $m=m_1$ и $m=m_2$ [3]:

$$J_{m_1}(W,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(w_j(x_i)\right)^{m_1}\cdot d_{ji}^2+\sum_{j=1}^{c}\eta_j^2\cdot\sum_{i=1}^{n}\left(1-w_j(x_i)\right)^{m_1}, \quad (3)$$

$$J_{m_2}(W,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(w_j(x_i)\right)^{m_2}\cdot d_{ji}^2+\sum_{j=1}^{c}\eta_j^2\cdot\sum_{i=1}^{n}\left(1-w_j(x_i)\right)^{m_2}. \quad (4)$$

Использование различных значений «ширины зоны» η_j ($j=\overline{1,c}$) определяет различные целевые функции в РСМ-алгоритме на основе ИНМТ2 при $\eta_j=\eta_{j1}$ и $\eta_j=\eta_{j2}$:

$$J_{\eta_{j1}}(W,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(w_j(x_i)\right)^m\cdot d_{ji}^2+\sum_{j=1}^{c}\eta_{j1}^2\cdot\sum_{i=1}^{n}\left(1-w_j(x_i)\right)^m, \quad (5)$$

$$J_{\eta_{j2}}(W,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}\left(w_j(x_i)\right)^m\cdot d_{ji}^2+\sum_{j=1}^{c}\eta_{j2}^2\cdot\sum_{i=1}^{n}\left(1-w_j(x_i)\right)^m. \quad (6)$$

Вычисление функции типичности $w_j(x_i)$ и координат центров кластеров реализуется с использованием итерационного алгоритма Карника – Менделя. Поиск оптимальных параметров РСМ-алгоритма на основе ИНМТ2, обеспечивающих адекватные результаты кластеризации, не возможен без применения какого-либо оптимизационного алгоритма, например, генетического алгоритма (ГА).

Генетический алгоритм поиска оптимальной комбинации значений фаззификаторов, реализующих управление неопределенностью, и значений «ширины зоны» для РСМ-алгоритма на основе интервальных нечетких множеств второго типа

Метод возможностной кластеризации с использованием РСМ-алгоритма на основе ИНМТ2 и ГА позволяет значительно сократить время поиска оптимальной комбинации значений параметров алгоритма кластеризации и обеспечить получение адекватных результатов кластеризации [3].

Пусть, например, в РСМ-алгоритме на основе ИНМТ2 для каждого кластера задается единственное значение «ширины зоны» η_j ($j=\overline{1,c}$), а для фаззификатора m определяется комбинация значений m_1 и m_2. В этом случае для поиска оптимальной комбинации значений фаззификаторов m_1, m_2 и оптимальных значений «ширины зоны» η_j ($j=\overline{1,c}$) может быть использован ГА.

При этом хромосома задается в виде

$$s=(m_1,m_2,\eta_1,\ldots,\eta_c), \quad (7)$$

где $m_1,m_2\in(1,\ m_{max}]$; m_{max} – некоторое действительное число, определяющее максимальное значение фаззификатора; $m_1<m_2$; η_j – «ширина зоны» j-го кластера ($j=\overline{1,c}$); $\eta_j\in[\eta_j^{min},\eta_j^{max}]$; $\eta_j^{\min}>0$, $\eta_j^{\max}>0$, $\eta_j^{min}<\eta_{max}$, $\eta_j^{max}\le\eta_{max}$, $\eta_j^{\min}<\eta_j^{\max}$, η_{max} – некоторое действительное число, определяющее максимальное значение «ширины зоны».

При этом длина хромосомы равна $2+c$.

В качестве функции соответствия для ГА может использоваться общий гиперобъем H [2, 3]:

$$H = \sum_{j=1}^{c} \left(\det\left(R_j\right)\right)^{\frac{1}{2}} \to \min , \quad (8)$$

$$R_j = \frac{1}{n_j} \cdot \sum_{i=1}^{n} \left(x_i - v_j\right) \cdot \left(x_i - v_j\right)^T , \quad (9)$$

где R_j – ковариационная матрица j-го кластера; n_j – количество объектов кластеризации, отнесенных к j-му кластеру; v_j – вектор координат центра j-го кластера; x_i – вектор координат (оценок по критериям) i-го объекта; $\det\left(R_j\right)$ – определитель ковариационной матрицы j-го кластера; n – количество объектов; c – количество кластеров; $\mathrm{i} = \overline{1,\mathrm{n}}$, $\mathrm{j} = \overline{1,\mathrm{c}}$.

При реализации ГА следует использовать одноточечное скрещивание, а количество мутирующих генов не должно превышать 10–20 % от длины хромосомы.

При выполнении операции скрещивания выбирается вероятность скрещивания R_c и генерируется случайное число N_c. Если $R_c > N_c$, то случайным образом выбирается точка скрещивания z и выполняется скрещивание. Если в качестве точки скрещивания выбирается второй ген, определяющий фаззификатор m_2, то скрещивание выполняется без проверки каких-либо условий. Если в качестве точки скрещивания выбирается первый ген, определяющий фаззификатор m_1, то при выполнении операции скрещивания для обоих хромосом-отпрысков выполняется проверка условия: $m_1 < m_2$. Если это условие не выполняется, то осуществляется выбор новой точки скрещивания до тех пор, пока в точке скрещивания для обоих хромосом-отпрысков не будет выполнено условие $m_1 < m_2$ либо в качестве точки скрещивания не будет выбран второй ген, определяющий фаззификатор m_2, или ген, соответствующий «ширине зоны» η_j j-го кластера ($j = \overline{1,c}$).

При выполнении операции мутации выбирается вероятность мутации R_m и генерируется случайное число N_m. Если $R_m > N_m$, то случайным образом выбирается точка мутации z и выполняется мутация. Если в качестве точки мутации выбирается первый или второй ген, то при выполнении мутации для хромосы-отпрыска выполняется проверка условия: $m_1 < m_2$. Если это условие не выполняется, то осуществляется выбор новой точки мутации до тех пор, пока в точке мутации для хромосомы-отпрыска не будет выполнено условие $m_1 < m_2$ либо в качестве точки мутации не будет выбран ген, соответствующий «ширине зоны» η_j j-го кластера ($j = \overline{1,c}$).

Тогда генетический алгоритм имеет вид.

1. Случайным образом создается популяция размером P. При этом для первого и второго генов выполняется проверка условия: $m_1 < m_2$.

2. При $g < G$ (G и g – максимальное и текущее количество поколений ГА соответственно) реализуется РСМ-алгоритм на основе ИНМТ2 с вычислением значения функции соответствия для каждой хромосомы и создается $R_c \cdot P/2$ пар хромосом-родителей.

3. Выполняются операции скрещивания и мутации для текущей популяции. При этом для первого и второго генов выполняется проверка условия: $m_1 < m_2$. Для хромосом-отпрысков реализуется РСМ-алгоритм на основе ИНМТ2 и вычисляются значения функции соответствия.

4. Создается новая популяция размером $\left(P + R_c \cdot P\right)$, дополненная хромосомами-отпрысками в количестве $R_c \cdot P$, затем $R_c \cdot P$ хромосом с худшими значениями функции соответствия отбрасываются. Если $g < G$, осуществляется переход к шагу 2.

5. Выбирается лучшая хромосома, которая минимизирует функцию соответствия. Для каждого объекта определяется его принадлежность к кластерам.

Одновременно с популяцией хромосом вида (7) существуют «популяции» значений функции соответствия, координат центров кластеров и степеней принадлежности объектов центрам кластеров.

В случае неопределенности выбора значений «ширины зоны» η_j ($j = \overline{1,c}$) для РСМ-алгоритма на основе ИНМТ2 соответствующий ГА поиска оптимальных параметров метода возможностной кластеризации реализуется аналогичным образом.

Приведенный выше метод кластеризации и соответствующий ему ГА могут быть модифицированы, если есть какие-либо веские соображения по выбору значений фаззификаторов m_1 и m_2. Например, в качестве значений фаззификаторов m_1 и m_2, реализующих управление неопределенностью, для РСМ-алгоритма на основе ИНМТ2 могут использоваться значения фаззификаторов m_1 и m_2, полученные с помощью FCM-алгоритма на основе ИНМТ2. В этом случае при инициализации РСМ-алгоритма на основе ИНМТ2 следует использовать координаты центров кластеров, вычисленные с помощью FCM-алгоритма на основе ИНМТ2. При этом в ряде случаев удается не только уменьшить размерность оптимизационной задачи, но и улучшить результаты кластеризации за счет учета свойства кластерной типичности.

При фиксированных значениях фаззификаторов m_1 и m_2 хромосома может быть представлена в виде

$$s = \left(\eta_1, ..., \eta_c\right), \tag{10}$$

где η_j – «ширина зоны» j-го кластера ($j = \overline{1,c}$); $\eta_j \in [\eta_j^{min}, \eta_j^{max}]$; $\eta_j^{\min} > 0$, $\eta_j^{\max} > 0$, $\eta_j^{min} < \eta_{max}$, $\eta_j^{max} \le \eta_{max}$, $\eta_j^{\min} < \eta_j^{\max}$, η_{max} – некоторое действительное число, определяющее максимальное значение «ширины зоны».

В этом случае длина хромосомы равна c .

Предлагаемый модифицированный ГА реализуется так же, как и приведенный выше ГА, за исключением тех шагов, где для генов, определяющих значения фаззификаторов m_1 и m_2, необходимо выполнять проверку условия $m_1 < m_2$, так как значения фаззификаторов фиксированы.

Экспериментальные результаты

Ниже приведен пример кластеризации множества объектов на три кластера с использованием алгоритма четких c-средних и методов кластеризации на основе НМТ1 и ИНМТ2 для множества объектов, содержащего кластеры существенно разной плотности или существенно разного объема.

На рис. 1 показано множество объектов, содержащее три кластера существенно разного объема (объекты разных кластеров помечены маркерами разной формы). Кластеры представляют собой множества объектов, координаты которых были сгенерированы с использованием нормального закона распределения с центрами (10, 50), (50, 50) и (90, 50). При этом при генерации первой и второй координат объектов первого и третьего кластеров использовались нормальные законы распределения, имеющие одинаковые дисперсии. Координаты объектов второго кластера имеют существенно большую дисперсию по обеим координатам. Первый и второй кластеры содержат по 35 объектов, а второй кластер – 130 объектов.

На рис. 2–5 приведены результаты кластеризации с использованием различных алгоритмов кластеризации, при этом центры кластеров отмечены «белыми» маркерами треугольной формы.

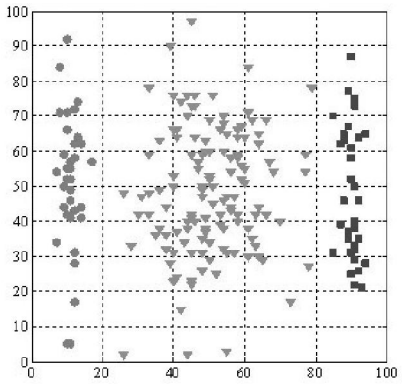

Рис. 1. Исходное множество объектов кластеризации

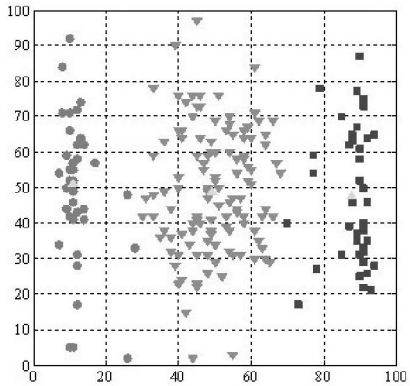

Рис. 2. Результаты кластеризации с использованием алгоритма четких c-средних

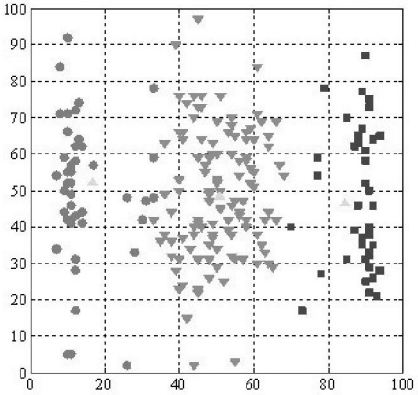

Рис. 3. Результаты кластеризации с использованием FCM-алгоритма на основе НМТ1 при $m = 2$

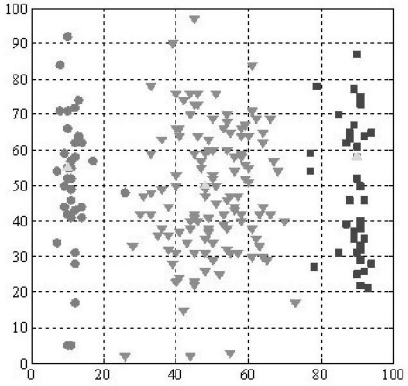

Рис. 4. Результаты кластеризации для FCM-алгоритма на основе ИНМТ2 для комбинации $m_1 = 108{,}813$ и $m_2 = 112{,}174$

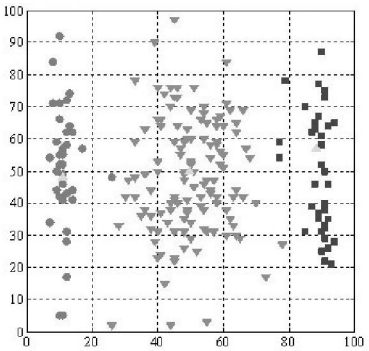

Рис. 5. Результаты кластеризации для PCM-алгоритма на основе ИНМТ2 для комбинации $m_1 = 108{,}813$, $m_2 = 112{,}174$, $\eta_1 = 115{,}410$, $\eta_2 = 148{,}455$ и $\eta_3 = 117{,}808$

В таблице приведены результаты кластеризации с использованием алгоритмов кластеризации на основе НМТ1 и ИНМТ2, подтверждающие эффективность использования алгоритмов кластеризации на основе ИНМТ2 для множеств объектов существенно разной плотности или существенно разного объема. Следует отметить, что для трех сгенерированных исходных кластеров (см. рис. 1) общий гиперобъем H составляет 279,725. Таким образом, при использовании алгоритмов кластеризации на основе ИНМТ2 удалось уменьшить ошибку кластеризации с минимального количества в 9 объектов (4,5 % от мощности множества объектов) для алгоритма четких c-средних до

4-5 объектов (2–2,5 % от мощности множества объектов) для алгоритмов кластеризации на основе ИНМТ2.

Результаты кластеризации

Алгоритм кластеризации	Значение функции соответствия	Ошибочно классифицированные объекты
Алгоритм четких *c*-средних	350,616	9 объектов (4,5 %)
FCM на основе НМТ1 при $m = 2$ (центры кластеров: первый кластер: (11,014, 54,997); второй кластер: (49,771, 50,962); третий кластер: (89,982, 57,994))	(нечеткий общий гиперобъем) 541,268	13 объектов (6,5 %)
FCM на основе ИНМТ2 для комбинации $m_1 = 108{,}813$ и $m_2 = 112{,}174$ (центры кластеров: первый кластер: (10,134, 54,962); второй кластер: (48,000, 49,999); третий кластер: (89,999, 57,999))	330,144	5 объектов (2,5 %)
РСМ на основе ИНМТ2, при фиксированных значениях фаззификаторов m_1 и m_2, определенных с помощью FCM-алгоритма (центры кластеров: первый кластер: (10,923, 48,029); второй кластер: (49,728, 49,883); третий кластер: (88,453, 56,894))	311,776	4 объекта (2 %)

Как видно из таблицы, лучший результат кластеризации (при минимальном значении общего гиперобъема H, равном 311,776) с ошибкой кластеризации в 4 объекта (2 % от мощности множества объектов) был получен с помощью РСМ-алгоритма на основе ИНМТ2 при фиксированных значениях фаззификаторов $m_1 = 108{,}813$ и $m_2 = 112{,}174$, определенных с помощью FCM-алгоритма на основе ИНМТ2.

Заключение

В статье рассмотрена проблема управления неопределенностью выбора параметров РСМ-алгоритма на основе ИНМТ2 при кластеризации множества объектов, содержащего кластеры существенно разной плотности или существенно разного объема.

Предложен метод кластеризации, разработанный для РСМ-алгоритма на основе ИНМТ2, позволяющий ослабить свойство кластерной относительности и учесть свойство кластерной типичности объектов, что в ряде случаев приводит к улучшению результатов кластеризации.

Применение ГА позволяет найти оптимальную комбинацию значений фаззификатора и «ширины зоны» в РСМ-алгоритме на основе ИНМТ2, обес-

печивающую лучшие результаты кластеризации, что подтверждается минимальным значением функции соответствия.

На практике предлагаемый метод кластеризации может быть применен, например, для решения задачи классификации технического состояния зданий и сооружений.

Библиографические ссылки

1. *Rhee F. C.-H.* Uncertain fuzzy clustering: insights and recommendations // IEEE Computational intellegence magazine, 2007. V. 2. № 1.

2. *Демидова Л. А., Кираковский В. В.* Кластеризация объектов на основе нечетких множеств второго типа и генетического алгоритма // Управление созданием и развитием систем, сетей и устройств телекоммуникаций / под ред. А. В. Бабкина, В. А. Кежаева: труды международной конференции. СПб., 2008.

3. *Демидова Л. А., Кираковский В. В.* Методы кластеризации объектов на основе нечетких множеств второго типа и генетического алгоритма // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2008. № 6(69).