УДК 004.048:519.816

Н.А. Новоселова, И.Э. Том

ЭВОЛЮЦИОННЫЙ МЕТОД НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

Предлагается эволюционный метод нечеткой кластеризации, использующий генетический алгоритм с переменной длиной хромосом, который позволяет находить близкое к оптимальному разбиение объектов на кластеры с одновременным определением их числа. Проводится теоретический анализ вычислительной сложности предложенного метода в сравнении со стандартным подходом к поиску количества кластеров в данных. При проведении тестирования метода на двух наборах данных показывается, что классификационные правила, построенные на основе предложенного метода кластеризации, имеют более высокую точность классификации, чем полученные классическим FCM-методом.

Введение

Задача кластеризации состоит в разбиении объектов данных на несколько подмножеств (кластеров) объектов, схожих между собой в некотором признаковом пространстве. В метрическом пространстве «схожесть» обычно определяют через расстояние. Расстояние может рассчитываться как между исходными объектами, так и между этими объектами и прототипами кластеров. Обычно координаты прототипов заранее неизвестны: они находятся одновременно с разбиением объектов данных на кластеры.

Кластерные методы можно разделить на три большие группы [1]: относящие объекты данных к частично пересекающимся кластерам, разделяющие и иерархические. Последние две группы методов связаны между собой, так как иерархическая кластеризация является следствием разделяющих методов, каждый из которых представляет собой разделение набора данных на взаимоисключающие подмножества. В этом случае каждый объект данных относится только к одному из кластеров. На практике большинство наборов данных включают подмножества, которые не могут быть строго разделены на непересекающиеся группы. Например, существуют ситуации, когда структура данных описывается категориями, которые в некоторой степени перекрываются. В этом случае рекомендуется использовать кластерные алгоритмы, которые способны работать с пересекающимися кластерами данных. Методы нечеткой кластеризации относятся к первой группе методов и позволяют находить нечеткие кластеры, содержащие объекты данных с частичной принадлежностью [2, 3]. Нечеткая кластеризация во многих ситуациях более «естественна», чем четкая, например для объектов, расположенных на границе кластеров. Кластерные методы широко используются при анализе многомерных данных для медицинских, финансовых, инженерных и других приложений, где необходимым является выявление групп объектов с похожими признаками и поведением, и на этой основе принимается решение в отношении нового объекта данных путем расчета его степени близости к той или иной первоначально определенной группе.

Одной из основных проблем при решении задачи кластеризации является определение количества кластеров, которое задает дальнейший ход разбиения объектов на группы и, как правило, является изначально неизвестным. Для преодоления этого недостатка широко используется подход, который заключается в многократном итерационном выполнении кластерного алгоритма для различного априори заданного числа кластеров и последующем выборе того количества кластеров, которое обеспечивает наилучший результат согласно некоторому показателю качества разбиения [3]. Такая последовательно выполняемая процедура кластеризации с увеличивающимся числом кластеров может быть эффективной только в случае малого количества кластеров в структуре данных, что не всегда имеет место на практике. При анализе набора объектов данных со сложной структурой хорошей альтернативой является применение эволюционного алгоритма для решения задачи кластеризации, так как такой алгоритм позволяет получить близкие к оптимальным решения за приемлемое время. В настоящей статье описывается эволюционный метод нечеткой кластеризации, который позволяет автоматически определять количество кластеров в наборе данных в процессе их кластеризации. Кроме того, следует отметить, что результаты тестирования предлагаемого метода на различных наборах данных пока-

зали его большую эффективность в сравнении с итерационными алгоритмами поиска количества кластеров в данных.

1. Задача нечеткой кластеризации и показатели качества разбиения

Решением задачи нечеткой кластеризации является разбиение $P = \{A^1, ..., A^c\}$ объектов из множества $X = \{x_1, \ldots, x_n\}$ на заданное число нечетких кластеров *c*. Нечеткие кластеры описываются матрицей нечеткого разбиения

$$U = [\mu_{ij}], \quad \mu_{ij} \in [0,1], \quad i = \overline{1,c},\ j = \overline{1,n}, \tag{1}$$

в которой *i*-й столбец содержит степени принадлежности объекта $\mathrm{x}_j = (x_j^1, \ldots, x_j^m)$, заданного вектором значений признаков, к кластерам $A^1, ..., A^c$. При этом должны выполняться следующие условия:

$$\sum_{i=1}^{c} \mu_{ij} = 1,\ j = \overline{1,n}; \quad 0 < \sum_{j=1}^{n} \mu_{ij} < n. \tag{2}$$

Задача нечеткой кластеризации заключается в нахождении экстремального значения некоторого функционала $F(P)$ на множестве всех нечетких разбиений:

$$F(P) \to \underset{P \in \Omega}{\mathrm{extr}}, \tag{3}$$

где Ω – множество всех возможных нечетких разбиений.

В литературе предложены самые разнообразные функционалы $F(P)$ и соответствующие алгоритмы их оптимизации. Функционал Беждека – Данна – наиболее изученный и распространенный вариант постановки задачи нечеткой кластеризации [4]:

$$F(P) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{\gamma} d(\mathrm{x}_j, \tau_i), \tag{4}$$

где γ – экспоненциальный вес (показатель нечеткости), $1 < \gamma < \infty$; $d(\mathrm{x}_j, \tau_i)$ – функция расстояния между объектами данных и центрами кластеров, в качестве которой, как правило, используется квадрат евклидовой нормы:

$$d(\mathrm{x}_j, \tau_i) = \left\| \mathrm{x}_j - \tau_i \right\|^2. \tag{5}$$

Тогда с учетом (5) выражение (4) можно записать как

$$F(P) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{\gamma} \left\| \mathrm{x}_j - \tau_i \right\|^2 \ [2]. \tag{6}$$

При решении задачи нечеткой кластеризации требуется определить такое оптимальное нечеткое c-разбиение $P = \{A^1, ..., A^c\}$, для которого функционал (4) принимает наименьшее значение.

В работе [3] предложено множество алгоритмов нечеткой кластеризации, основанных на минимизации критерия (4). Нахождение матрицы нечеткого разбиения *U* с минимальным значением критерия (4) представляет собой задачу нелинейной оптимизации, которая может быть

решена различными методами. Наиболее известный и часто применяемый метод решения этой задачи – алгоритм нечетких *C*-средних (*FCM*) [2], в основу которого положен метод неопределенных множителей Лагранжа. Он позволяет найти только локальный оптимум, поэтому выполнение алгоритма из различных начальных точек может привести к разным результатам.

Задача оптимизации функционала (4) может быть решена методом последовательных приближений в соответствии с алгоритмом *FCM* [5].

Как было упомянуто во введении, одной из главных проблем при использовании оптимизационных методов кластеризации, в том числе метода *C*-средних, является определение «реального» числа нечетких кластеров, на которые разбивается исследуемая совокупность данных. Для решения этой проблемы разными авторами были предложены различные показатели, характеризующие нечеткое разбиение. Количество кластеров обычно оценивается с помощью показателей качества разбиения, однако вопрос о том, как формально и достоверно определить правильность выбора количества кластеров для произвольного набора данных, все еще остается открытым.

Основные показатели, характеризующие качество разбиения данных на группы нечеткими алгоритмами кластеризации, приведены в работе [3].

Большинство исследователей рассматривает показатель качества разбиения как отображение Q множества U матриц разбиения на область действительных чисел: $Q : U \to R$. Экстремальное значение функции Q задает оптимальное количество кластеров или групп данных.

Для поиска экстремума показателя качества разбиения классические алгоритмы нечеткой кластеризации выполняются несколько раз, каждый раз с новым задаваемым количеством кластеров. Локальный оптимум показателя качества разбиения помогает определить наиболее соответствующее анализируемым данным количество кластеров.

Наиболее часто используемыми функционалами качества кластеризации являются индекс разбиения [2], энтропия разбиения [2] и показатель Хи – Бени [6].

Индекс качества разбиения. Индекс разбиения для матрицы U, обозначаемый $P(U)$, определяется как среднее значение квадратов значений функции принадлежности, фигурирующих в матрице нечеткого разбиения

$$P(U) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^2 . \tag{7}$$

Если каждый объект данных относится (принадлежит) только к одному кластеру (в случае четкой кластеризации), то значение индекса максимально и равно 1. Если объекты данных в равной мере принадлежат всем кластерам и значение принадлежности каждому из них равно $1/c$, то значение индекса $P(U)$ минимально и равно $1/c$. Чем больше значение, тем лучше качество разбиения объектов данных на кластеры, т. е. тем лучше выделяются отдельные кластеры данных.

Энтропия разбиения. Значение энтропии разбиения рассчитывается следующим образом:

$$H(U) = -\frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij} \ln(\mu_{ij}) . \tag{8}$$

Значения энтропии разбиения лежат в диапазоне $\left[0, \ln(c)\right]$, причем максимальное значение достигается при равномерном распределении значений принадлежностей объекта кластерам, равном $1/c$:

$$H(U) = -\frac{1}{n} \sum_{i=1}^{c} \sum_{j}^{n} \frac{1}{c} \ln(\frac{1}{c}) = \ln(c) .$$

Несмотря на то что описанные выше показатели качества разбиения пригодны для оценки матрицы разбиения, их использование для определения числа кластеров не является очевидным. К недостаткам этих показателей можно отнести:

1) монотонную зависимость от числа кластеров;

2) чувствительность к значению показателя нечеткости γ. При $\gamma \to 1$ значения показателей одинаковы для всех значений *c*, при $\gamma \to \infty$ оба показателя в качестве наилучшего разбиения определяют разбиение на два кластера;

3) отсутствие прямой связи с геометрией данных, т. е. с распределением объектов данных в многомерном признаковом пространстве, так как сами данные не используются при расчетах показателей качества разбиения.

К показателям, которые учитывают как значения принадлежностей, так и сами данные, относятся показатели Хи – Бени, Фукуяма – Сугено, а также показатели, предложенные в работах [3, 7]. В нашем исследовании для расчета качества разбиения объектов данных на кластеры используется показатель Хи – Бени.

Показатель Хи – Бени называют функцией оценки компактности и разделимости кластеров. Для нечеткого разбиения $U = [\mu_{ij}]$ нечеткое отклонение объекта x_j от центра кластера *i* определяется следующим образом:

$$d_{ij} = \mu_{ij} \left\| x_j - \tau_i \right\|.$$

Отклонение σ_i для кластера *i* определяется как сумма квадратов нечетких отклонений объектов из *X*. Общее отклонение кластеров σ равно сумме отклонений всех кластеров σ_i. Величина $\pi = \left(\sigma_i / n_i\right)$ обозначает компактность кластера *i*. Так как n_i равно количеству объектов, относящихся к кластеру *i*, то π – средняя вариация в кластере *i*.

Значение разделимости нечетких кластеров определяется как минимум расстояний между центрами кластеров:

$$d_{\min} = \min \left\| \tau_i - \tau_j \right\|.$$

Таким образом, показатель Хи – Бени определяется следующим образом:

$$XB = \pi / (n \cdot d_{\min}),$$

где *n* – количество объектов данных.

Малые значения показателя Хи – Бени соответствуют компактным и хорошо разделимым кластерам, однако значение показателя монотонно убывает с ростом количества кластеров. Для того чтобы исключить это влияние на оценку оптимального количества кластеров, для анализируемого набора данных обычно определяется граница количества кластеров $c_{\max}$ и поиск минимального значения показателя Хи – Бени осуществляется в диапазоне $[2, c_{\max}]$ путем многократного выполнения алгоритмов кластеризации. Общая схема стандартного подхода к поиску количества кластеров с использованием алгоритма *FCM* следующая:

1. *Выбрать значения c_{max} – максимальное количество кластеров, S_c – критерий останова, задать начальное значение показателя качества разбиения XB^*_{VC}.*

2. *Для c=2, …, c_{max} выполнить:*

{

Для i = 1, …,n_p выполнить:

{

2.1. Случайным образом генерировать разбиение с количеством кластеров c.

2.2. Запустить FCM до выполнения S_c.

2.3. Вычислить XB_{VC} для полученного разбиения.

2.4. Если ($XB_{VC} < XB^*_{VC}$), то

{ $XB^*_{VC} = XB_{VC}$.

$c^* = c$.

Сохранить полученное разбиение U с количеством кластеров c^*.

}

} // Конец цикла по i

} // Конец цикла по c

3. *Вывести {XB^*_{VC}, c^*, соответствующую матрице нечеткого разбиения U для c^*}.*

Представленный выше подход к оценке количества кластеров с использованием показателя качества разбиения объектов данных имеет существенный недостаток: требует многократного решения задачи кластеризации, что является неэффективным для многих практических применений.

Для того чтобы решить задачу определения числа кластеров за один прогон кластерного алгоритма, авторами разработан метод нечеткой кластеризации, использующий поисковые возможности генетического алгоритма (ГА) оптимизации. ГА в процессе эволюции позволяет одновременно с поиском решения оптимизационной задачи определить оптимальное количество кластеров в данных [8]. Для этого в качестве оптимизационного критерия был выбран показатель качества разбиения Хи – Бени.

2. Описание метода нечеткой кластеризации с использованием ГА

В связи с тем что нечеткая кластеризация может рассматриваться как задача нелинейной оптимизации, в настоящей работе описано применение ГА для поиска кластеров в данных. В литературе были предложены различные эволюционные алгоритмы для решения задач кластеризации. В работах [9, 10] генетические алгоритмы с различными схемами кодирования были использованы для решения задачи четкой кластеризации. Для поиска перекрывающихся кластеров эволюционный подход был использован в работах [11, 12]. Однако во всех перечисленных подходах необходимым являлось предварительное задание числа кластеров в данных. Предлагаемый метод нечеткой кластеризации на основе ГА использует особи с переменной длиной, что позволяет найти результат кластеризации, являющийся наилучшим относительно показателя качества разбиения Хи – Бени. Преимуществом предложенного метода является возможность за один прогон ГА найти не только оптимальное разбиение объектов данных на кластеры, но и определить их количество, которое соответствует минимуму показателя качества разбиения Хи – Бени.

Обозначим множество всех возможных нечетких разбиений через U:

$$U = \left\{ F \in R^{c \times n} \mid \sum_{i=1}^{c} \mu_{ij} = 1,\ 0 < \sum_{j=1}^{n} \mu_{ij} < n,\ \mu_{ij} \in [0,1] \right\}.$$

Наилучшая матрица нечеткого разбиения F^* определяется следующим образом:

$$F^* \in U, \quad XB\left(F^*, V^*, X\right) = \min_{F^* \in U} XB\left(F, V, X\right),$$

где XB – значение показателя Хи – Бени; V – множество центров кластеров; X – множество объектов данных.

Количество кластеров и соответствующие значения координат центров кластеров эволюционируют одновременно в процессе работы ГА. В связи с тем что длина особи является переменной, в одной популяции могут содержаться особи, кодирующие разное количество кластерных центров.

Отдельные особи популяции представляют собой решение задачи кластеризации и кодируются действительными числами, задающими координаты центров кластеров $\tau_i \in V$, $i = 1, \ldots, c$. Длина i-й особи $L_i = c \cdot m$, где c – количество кластеров; m – размерность объектов данных или количество признаков. Например, для трехмерных объектов данных особь (1,1; 4; 6,3)(7,5; 2,4; 3,8) представляет собой координаты двух кластерных центров. Каждая особь популяции изначально кодирует центры $c \in [2, c_{\max}]$ кластеров, где c_{max} – предварительно заданное максимальное количество кластеров в особи начальной популяции. В процессе эволюции длина особи может варьироваться произвольным образом и не ограничена в дальнейшем значением c_{max}. Начальная популяция генерируется случайным образом с использо-

ванием объектов данных. Центры кластеров τ_i, закодированные в особи, представляют собой координаты объектов данных, которые выбраны из набора данных случайным образом.

Значение оценочной функции показывает степень приспособленности особи и эффективность решения оптимизационной задачи. В качестве оценочной функции используется значение, обратное показателю нечеткого разбиения Хи – Бени. Показатель Хи – Бени определяется как отношение общей вариации σ значений объектов данных к минимуму *div* значений разделимости кластеров следующим образом:

$$\sigma(U,V,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2 d(\mathrm{x}_j,\tau_i^k);$$

$$div(V)=\min_{i\neq j}\{\left\|\tau_i-\tau_j\right\|^2\};$$

$$XB(U,V,X)=\frac{\sigma(U,V,X)}{n\cdot div(V)},$$

где c – количество кластеров; n – количество объектов данных, $d(\mathrm{x}_j,\tau_i)=\left\|\mathrm{x}_j-\tau_i\right\|^2$. Оценочная функция рассчитывается как $f=1/XB(U,V,X)$. Наилучшее нечеткое разбиение соответствует наибольшему значению функции f, при этом значение σ должно быть малым, в то время как значение разделимости центров кластеров *div* – достаточно большим.

Формирование популяции особей новой генерации выполняется следующим образом:

1. Из каждой особи декодируются координаты кластерных центров, а затем рассчитываются значения принадлежности μ_{ij} каждого объекта данных каждому из кластеров:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d(\mathrm{x}_j,\tau_i)}{d(\mathrm{x}_j,\tau_k)}\right)^{\frac{1}{\gamma-1}}},\quad 1\le i\le c,\ 1\le j\le n.$$

Если при этом значение $d(\mathrm{x}_j,\tau_i)$ для некоторого центра кластера τ_i равно нулю, то $\mu_{ij}=1$ для $k=i$ и $\mu_{ij}=0$ для $k=1,\ldots,c$ и $k\neq i$.

2. Значения центров кластеров обновляются согласно формуле

$$\tau_k=\frac{\sum_{j=1}^{n}(\mu_{kj})^{\gamma}\mathrm{x}_j}{\sum_{j=1}^{n}(\mu_{kj})^{\gamma}},\quad 1\le k\le c.$$

3. Для каждой особи рассчитывается значение оценочной функции f.

4. Выполняются операции отбора особей, для чего применяется обычный пропорциональный отбор. Каждая особь копируется в новую популяцию количество раз, пропорциональное значению ее функции оценки, т. е. более приспособленные особи выживают и копируются в новую популяцию.

5. Выполняется операция рекомбинации (кроссовер и мутация) для модификации отдельных особей популяции.

При кроссовере (скрещивании) родительских особей популяции кластерные центры считаются неделимыми, т. е. точка скрещивания может лежать только между координатами двух соседних кластерных центров, причем при скрещивании гарантируется, что дочерние особи кодируют не менее двух кластерных центров. Операция скрещивания реализуется следующим образом:

1) предположим, что две родительские особи i_1 и i_2 состоят из M_1 и M_2 кластеров соответственно;

2) определим точку скрещивания для родительской особи i_1 как $\tau_1=rand()\bmod M_1$, т. е. τ_1 может принимать значения от 0 до M_1–1. Для того чтобы дочерние особи кодировали не менее

двух кластеров, точка скрещивания τ_2 для второй особи i_2 должна находиться в некотором диапазоне $[L_{bound}, U_{bound}]$, где L_{bound} и U_{bound} рассчитываются по формулам

$$L_{bound} = \min\left(2, \max\left(0, 2-\left(M_1 - \tau_1\right)\right)\right) \quad \text{è} \quad U_{bound} = \left(M_2 - \max\left(0, 2-\tau_1\right)\right);$$

3) точка скрещивания для второй особи определяется как

$$\tau_2 = \begin{cases} L_{bound} + rand() \bmod\left(U_{bound} - L_{bound} + 1\right), \text{åñëè } L_{bound} \le U_{bound}; \\ 0, \text{åñëè } L_{bound} > U_{bound}. \end{cases}$$

После выполнения с вероятностью P_{cross} операции кроссовера с вероятностью P_{mute} выполняется операция мутации отдельных генов (координат кластеров) новых особей. Для этого первоначально генерируется случайное число τ, равномерно распределенное в диапазоне $\tau \in [0,1]$. Новое значение координаты, закодированной в гене, рассчитывается следующим образом:

$$v^* = \begin{cases} (1 \pm 2 \cdot \tau) \cdot v, \text{åñëè } v \neq 0; \\ \pm 2 \cdot \tau, \text{ åñëè } v = 0. \end{cases}$$

Использование арифметических операций ‘+’ или ‘–‘ равновероятно. С учетом вероятности P_{mute} выполнения операции мутации значения координат могут быть изменены более чем у одного кластерного центра.

6. Наилучшая особь предыдущей популяции копируется в новую популяцию без изменений, что соответствует операции элитаризма ГА.

Все перечисленные выше шаги реализуются конечное число раз до выполнения условия останова ГА. Таковым условием является либо достижение максимального количества генераций, либо отсутствие изменений координат кластеров в лучшей относительно значения оценочной функции особи.

По окончании работы ГА, лежащего в основе предлагаемого метода нечеткой кластеризации, получаем особь, содержащую $c \cdot m$ генов, где c – определенное количество кластеров, m – размерность признакового пространства. Границы соответствующего кластера определяются из матрицы нечеткого разбиения объектов данных по значению принадлежности к тому или иному кластеру. В случае принадлежности объекта нескольким кластерам выбор осуществляется по максимальному значению.

Как правило, решение задачи кластеризации не является конечной целью исследования набора данных, а предваряет задачу классификации объектов данных. В нашем случае результаты нечеткой кластеризации были использованы для построения набора нечетких классификационных правил [8]. Функции принадлежности предпосылок и следствия правил определяются с помощью полученного нечеткого разбиения объектов данных в m-мерном пространстве признаков. Каждый из полученных кластеров определяет одно нечеткое правило. С математической точки зрения степень принадлежности значения признака y p-й проекции $\mu_k^{(p)}$, $p \in [1,m]$, нечеткого кластера k равна супремуму по всем степеням принадлежности всех объектов данных кластеру k, p-я компонента которых равна y:

$$\mu_k^{(p)}(y) = \sup\left\{ \frac{1}{\sum_{j=1}^{c}\left(\frac{d^2(\tau_k, \mathbf{x})}{d^2(\tau_j, \mathbf{x})}\right)^{\frac{1}{\gamma-1}}} \middle| \mathbf{x} = \left(x_1, \ldots, x_{p-1}, y, x_{p+1}, \ldots, x_m\right) \in R^m \right\}. \tag{9}$$

Так как расчет всех степеней принадлежности по формуле (9) является достаточно сложным, используется упрощенная процедура. Согласно этой процедуре функции принадлежности

предпосылки генерируются путем поточечного проецирования матрицы нечеткого разбиения на одномерные координатные пространства признаков [13], результатом чего являются одномерные дискретные нечеткие множества. Для преобразования дискретной функции принадлежности в непрерывную рассчитывается ее выпуклая оболочка, которая в дальнейшем аппроксимируется треугольной или трапециевидной функцией с помощью эвристического алгоритма, минимизирующего сумму квадратов ошибок [14].

3. Проверка эффективности классификационных правил, построенных на основе разработанного метода кластеризации

Нечеткие правила представляются в следующем виде:

R_i: Если $x \in A$, то класс C_1 с весом p_{i1} и … и класс C_M с весом p_{iM}, где $i = 1, \ldots, c$; A – нечеткое множество, заданное многомерной функцией принадлежности, соответствующей нечеткому кластеру; $\mathbf{x} = (x^1, ..., x^m)$ – m-мерный объект данных; M – количество меток класса; C_j – метка класса в следствии правила ($j = 1, \ldots, M$); p_{ij} – степень достоверности правила R_i для класса C_j. Для определения метки класса для нового объекта данных с помощью набора нечетких правил используется следующий алгоритм рассуждений.

Алгоритм 1

1. Рассчитывается степень принадлежности μ_{ik} объекта данных x_k каждому нечеткому кластеру i ($i = 1, \ldots, c$) с помощью выражения

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{\gamma - 1}}}. \tag{10}$$

2. С целью определения для объекта данных x_k «голоса» для каждого класса или степени отнесения объекта данных к каждому из имеющихся классов [7] используется выражение

$$V_{C_j}(x_k) = \sum_{R_i,\, i=1\ldots,c} \mu_{ik} \cdot p_{ij}, \quad j = 1, \ldots, M. \tag{11}$$

3. Метка класса для объекта данных x_k определяется в соответствии с меткой класса, который имеет большинство «голосов»: $x_k \to C_{j^*}$, где

$$V_{C_{j^*}} = \max \{ V_{C_j}(x_k) \mid j = 1, \ldots, M \}. \tag{12}$$

Степени достоверности p_{ij} для каждого класса C_j ($j = 1, \ldots, M$) и для каждого правила R_i ($i = 1, \ldots, c$) определяются на основе результатов нечеткой кластеризации набора данных $X = \{x_1, x_2, \ldots, x_n\}$ следующим алгоритмом.

Алгоритм 2

1. Степени принадлежности μ_{ik} всех объектов данных x_k ($k = 1, \ldots, n$) к каждому из кластеров i ($i = 1, \ldots, c$) рассчитываются с помощью выражения (10).

2. Для каждого класса C_j ($i = 1, \ldots, M$) рассчитывается сумма β_j степеней принадлежности объектов данных нечеткому правилу R_i ($i = 1, \ldots, c$):

$$\beta_j(R_i) = \sum_{x_k \in C_j} \mu_{ik}, \quad j = 1, \ldots, M. \tag{13}$$

3. Степень достоверности p_{ij} для каждого класса C_j ($j = 1, \ldots, M$) и для каждого правила R_i ($i = 1, \ldots, c$) рассчитывается следующим образом:

$$p_{ij} = \frac{\beta_j(R_i)}{\sum_{k=1}^{M} \beta_k(R_i)}, i = 1,\ldots,c,\ j = 1,\ldots,M. \quad (14)$$

Предложенный метод нечеткой кластеризации с использованием ГА и последующей процедурой построения набора нечетких правил классификации протестирован на искусственно сгенерированном наборе данных set_all и на наборе данных Iris из международного архива данных по машинному обучению.

Набор данных set_all (рис. 1) состоит из 330 объектов данных, характеризующихся двумя непрерывными признаками $x = (x_1, x_2)$, и разбит на три группы, сгенерированные случайным образом из нормального распределения. При этом каждая из групп соответствует различным средним значениям $v = (v_1, v_2)$ и стандартным отклонениям $\sigma = (\sigma_1, \sigma_2)$ нормального распределения.

Согласно предложенному методу кластеризации при нескольких запусках ГА были определены четыре кластера, обеспечивающие минимальное значение показателя Хи – Бени. После построения нечетких правил на основе результатов кластеризации точность классификации объектов данных set_all составила 99,7 %, т. е. всего один объект данных был неправильно классифицирован. В случае использования *FCM*-алгоритма для кластеризации данных set_all и последующего построения набора нечетких правил четыре объекта данных были неправильно проклассифицированы. На рис. 2 показаны результаты кластеризации и классификации набора данных set_all с использованием предложенного метода.

Второй исследуемый набор данных Iris состоит из трех классов (Setosa, Versicolour, Virginica), каждый класс данных включает 50 объектов, которые характеризуются значениями четырех признаков $\mathbf{x} = (x_1, \ldots, x_4)$. Класс Setosa линейно отделим от двух других классов, в то время как классы Versicolour и Virginica не являются линейно отделимыми. Наилучшим результатом работы стандартного кластерного метода *FCM* согласно значению показателя качества Хи – Бени является разбиение на два кластера. *FCM*-алгоритм кластеризации был применен девять раз для каждого количества кластеров из диапазона $c = \overline{2,\ldots,10}$, после чего были построены правила классификации и осуществлена классификация данных Iris (рис. 3). Ошибка классификации составила 10,7 %.

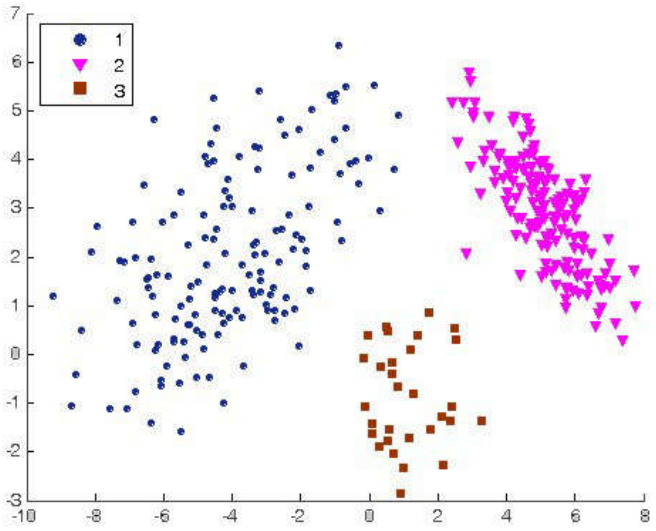
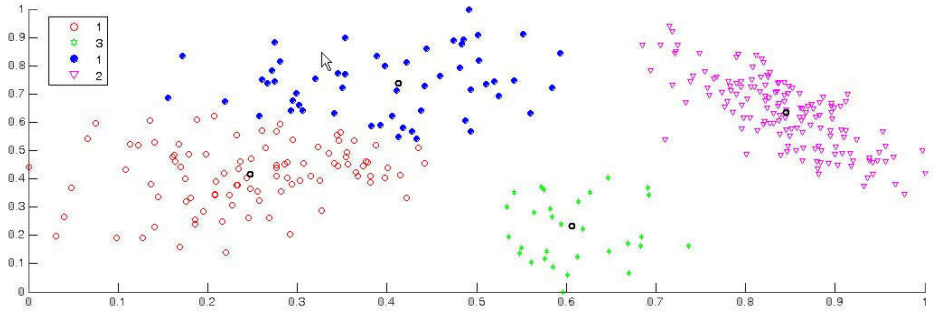
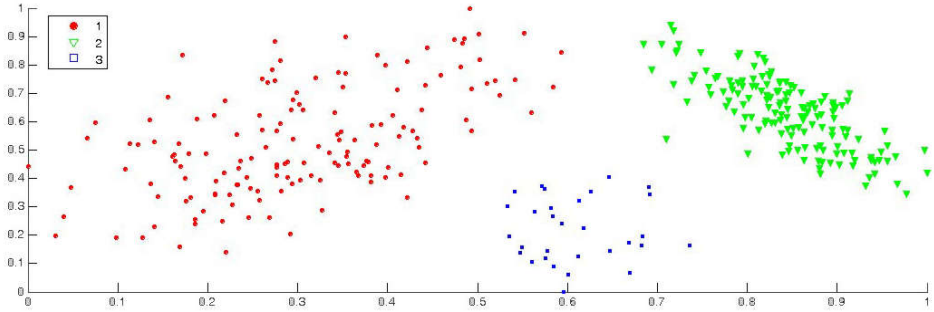


Рис. 1. Искусственно сгенерированный набор данных set_all

При использовании предложенного в данной работе кластерного алгоритма в результате восьми из десяти прогонов ГА данные были разделены на три кластера со значением показателя Хи – Бени, равным 0,15. При этом ошибка классификации равнялась 6,6 %, что на 4,1 % лучше, чем результаты, полученные *FCM*-алгоритмом.



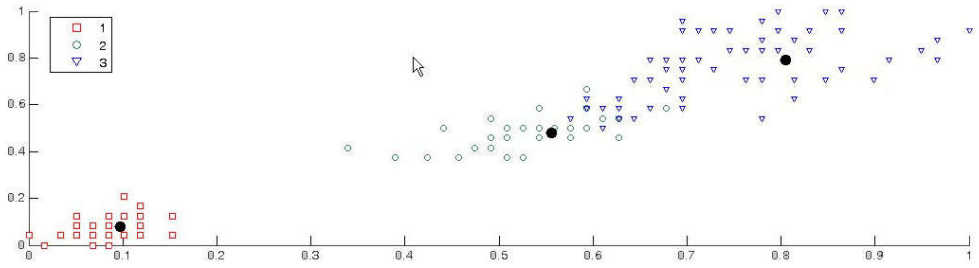
а)



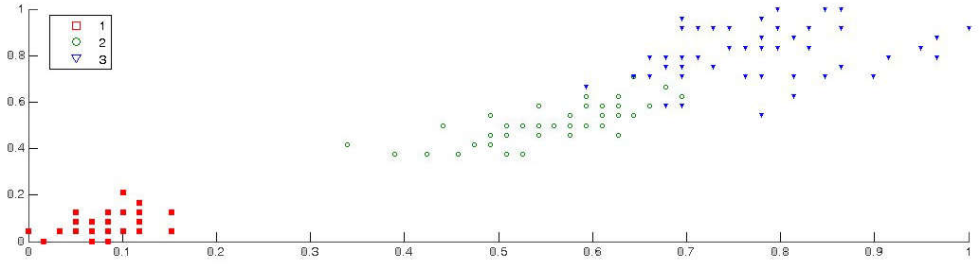
б)

Рис. 2. Результаты анализа набора данных set_all: *а)* кластеризация; *б)* классификация

Таким образом, предложенный метод нечеткой кластеризации позволяет не только автоматически определить соответствующее данным количество кластеров, но и в сравнении с классическим алгоритмом кластеризации *FCM* повысить точность классификации данных.



а)



б)

Рис. 3. Результаты анализа набора данных Iris в признаковом пространстве x_3–x_4:
а) кластеризация; *б)* классификация

4. Анализ вычислительной сложности

В настоящей работе выполнен сравнительный теоретический анализ вычислительной сложности предложенного эволюционного метода нечеткой кластеризации и стандартного подхода к определению количества кластеров данных с многократным запуском кластерного алгоритма.

В стандартном подходе, схема которого приведена в разд. 1, вычислительная сложность в большой степени зависит от количества кластеров в наборе данных. Если априори известен диапазон, в котором расположено действительное количество кластеров, и это количество сравнительно небольшое, то применение стандартного подхода может быть достаточно эффективным. Если же нет никакой информации о количестве кластеров в данных или есть вероятность, что данные содержат большое количество кластеров, – стандартный подход неэффективен, так как при этом необходимо запускать кластерный алгоритм большое количество раз. При этом можно и не найти реальное число кластеров в данных, когда оно не входит в априори выбранный диапазон.

При оценке вычислительной сложности стандартного подхода предполагается отсутствие предварительных знаний о количестве кластеров. В этом случае вычислительная сложность стандартного подхода с многократным запуском *FCM*-алгоритма оценивается как $O\left(t \cdot n \cdot c^2 \cdot m\right)$, где t – количество итераций *FCM*; n – количество объектов данных; c – количество нечетких кластеров; m – количество признаков. Принимая во внимание, что согласно стандартному подходу алгоритм *FCM* запускается n_p раз для каждого числа кластеров c из диапазона $[2, c_{\max}]$, общие вычислительные затраты оцениваются как

$$O\left(n_p \cdot t \cdot n \cdot \left(2^2 + 3^2 + \ldots + c_{\max}^2\right) \cdot m\right) \Rightarrow O\left(n_p \cdot t \cdot n \cdot c_{\max}^3 \cdot m\right), \quad (15)$$

где n_p – количество разбиений, выполняемых для каждого числа кластеров $c \le c_{\max}$.

С учетом только наиболее критических переменных общую вычислительную сложность подхода можно оценить как $O\left(n \cdot c_{\max}^3\right)$. Общую вычислительную сложность предлагаемого эволюционного метода нечеткой кластеризации можно оценить как $O\left(G \cdot P_s \cdot n \cdot \tilde{c}_{\max}^2 \cdot m\right)$, где G – количество генераций; P_s – размер популяции; $\tilde{c}_{\max}^2$ – максимальное количество кластеров, которые могут быть закодированы в особи в процессе эволюционного поиска. Величины G и P_s обратно пропорциональны, так как увеличение размера популяции обычно сокращает количество генераций, необходимых для сходимости ГА. Произведение этих величин не зависит от количества объектов данных и максимального количества кластеров. Таким образом, общая вычислительная сложность с учетом наиболее критических переменных оценивается как $O\left(n \cdot \tilde{c}_{\max}^2\right)$.

В случае отсутствия предварительной информации о количестве кластеров в данных предложенный эволюционный метод теоретически является более быстрым в вычислительном плане. Для подтверждения этого вывода необходимо проведение дополнительных экспериментов с различными наборами данных и статистического анализа полученных результатов, что является направлением наших дальнейших исследований.

Заключение

Предложенный в работе новый метод нечеткой кластеризации отличается от своих аналогов:

– способом кодирования особей ГА, согласно которому кодируются координаты центров кластеров, представляемые действительными числами;

– специальной операцией скрещивания, гарантирующей, что каждый из потомков кодирует центры по крайней мере двух кластеров;

– оценкой функции приспособленности особи показателем качества разбиения Хи–Бени, который равен отношению значения общей вариации данных к минимальному значению разделимости кластеров.

Проведенный теоретический анализ вычислительной сложности предложенного метода в сравнении со стандартным подходом к поиску количества кластеров в данных показал, что в реальных задачах эволюционный метод может быть гораздо быстрее стандартного подхода, что, конечно, не исключает эффективности последнего в случаях, когда имеется предварительная информация о количестве кластеров в данных. Для подтверждения этого теоретического вывода необходимо проведение ряда дополнительных вычислительных экспериментов.

Предложенный метод нечеткой кластеризации в сочетании с ранее разработанной авторами процедурой построения набора нечетких правил классификации протестирован на наборах данных set_all и Iris из международного архива данных по машинному обучению. При классификации объектов данных set_all точность составила 99,7 %, т. е. всего один объект данных был неправильно классифицирован, а для *FCM*-алгоритма Беждека четыре объекта данных были классифицированы неправильно. Для набора данных Iris точность классификации предложенного метода и *FCM*-алгоритма Беждека составила 93,4 и 89,3 % соответственно.

Список литературы

1. Jain, A.K. Algorithms for clustering data / A.K. Jain, R.C. Dubes. – Englewood Cliffs : Prentice-Hall, 1988. – 334 p.

2. Bezdek, J.C. Pattern recognition with fuzzy objective function algorithms / J.C. Bezdek. – N.Y. : Kluwer Academic Publishers, 1981. – 272 p.

3. Fuzzy Cluster Analysis / F. Hoppner [et al.]. – John Wiley and Sons, 1999. – 289 p.

4. Bezdek, J.C. Statistical parameters of cluster validity functionals / J.C. Bezdek, M.P. Windham, R. Elrich // International Journal of Parallel Programming. – 1980. – Vol. 9, № 4. – P. 323–336.

5. Bezdek, J.C. Norm-induced shell-prototypes (NISP) clustering / J.C. Bezdek, R.J. Hathaway, N.R. Pal // Neural, parallel and scientific computations. – 1995. – № 3. – P. 431–450.

6. Xie, X.L. A Validity Measure for Fuzzy Clustering / X.L. Xie, G. Beni // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1991. – Vol. 13, № 4. – P. 841–846.

7. Smyth, P. Clustering using Monte Carlo Cross-Validation / P. Smyth // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. – CA, Menlo Park : AAAI Press, 1996. – P. 126–133.

8. Novoselova, N. Supervised fuzzy clustering using genetic algorithm for the fuzzy classifier construction / N. Novoselova // Искусственный интеллект. – 2007. – № 4. – С. 343–351.

9. Hruschka, E.R. A genetic algorithm for cluster analysis / E.R. Hruschka, N.F.F. Ebecken // Intelligent data analysis. – 2003. – Vol. 7, № 1. – P. 15–25.

10. Kuncheva, L.I. Selection of cluster prototypes from data by a genetic algorithm / L.I. Kuncheva, J.C. Bezdek // Proc. of 5th European Congress on Intelligent techniques and Soft computting. – Aachen, Germany, 1997. – P. 1683–1688.

11. Klawonn, F. Fuzzy clustering with evolutionary algorithms / F. Klawonn // Proc. of 7th IFSA World Congress. – Prague, Czech Republic, 1997. – P. 957–962.

12. Hall, L.O. Clustering with a genetically optimized approach / L.O. Hall, I.B. Ozyurt, J.C. Bezdek // IEEE Trans. Evol. Comput. – 1999. – Vol. 3, № 2. – P. 103–112.

13. Klawonn, F. Constructing a fuzzy controller from data / F. Klawonn, R. Kruse // Fuzzy Sets and Systems. – 1997. – Vol. 85, № 2. – P. 177–193.

14. Sugeno, M. A fuzzy-logic-based approach to qualitative modeling / M. Sugeno, T.Yasukawa // IEEE Transactions on Fuzzy Systems. – 1993. – Vol. 1, № 1. – P. 7–31.

Поступила 04.05.09

Объединенный институт проблем информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: novosel@newman.bas-net.by

N.A. Novoselova, I.E. Tom

EVOLUTIONARY METHOD OF FUZZY CLUSTERING

The evolutionary method of fuzzy clustering of multivariate data is proposed. The method makes use the genetic algorithm with chromosomes of variable length, which allows to find the near-optimal cluster partition and simultaneously to define the number of clusters. A theoretical analysis of computational complexity of the proposed method in comparison with standard approach to the definition of the number of clusters is conducted. The results of testing on two data sets have demonstrated a more accurate classification of data with the classification rules constructed on the basis of the proposed method when compared to the classical FCM-method.