УДК 510.22+519.237.8

МЕТОД ЧАСТИЧНОГО ОБУЧЕНИЯ ДЛЯ ЭВРИСТИЧЕСКОГО АЛГОРИТМА ВОЗМОЖНОСТНОЙ КЛАСТЕРИЗАЦИИ ПРИ НЕИЗВЕСТНОМ ЧИСЛЕ КЛАССОВ

Канд. филос. наук ВЯТЧЕНИН Д. А.

Объединенный институт проблем информатики НАН Беларуси

В задачах сегментации изображений, обработки результатов научных исследований, при проектировании разнообразных систем поддержки принятия решений особая роль отводится нечетким методам автоматической классификации, в специальной литературе [1] именуемым также методами нечеткой кластеризации или нечеткими методами численной таксономии. В задачах кластеризации данные об исследуемой совокупности традиционно представляются либо матрицей $X_{n\times m}=[\hat{x}_i^t]$, $i=1,\ldots,n$, $t=1,\ldots,m$, именуемой матрицей «объект–признак», где x_i, $i=1,\ldots,n$ – объекты исследуемой совокупности X, а $\hat{x}^t$, $t=1,\ldots,m$ – значения признаков объектов $x_i\in X$, каждый из которых, таким образом, представляет собой точку в m-мерном признаковом пространстве, либо матрицей $\mathrm{P}_{n\times n}=[\rho_{ij}]$, $i,j=1,\ldots,n$ попарных коэффициентов близости или различия объектов, носящей название «объект–объект». При обработке данных методами нечеткой кластеризации результатом классификации является не только отнесение i-го объекта исследуемой совокупности $X=\{x_1,\ldots,x_n\}$ к l-му классу A^l, $l=1,\ldots,c$, но и указание функции принадлежности $u_{li}\in[0,1]$, $l=1,\ldots,c$, $i=1,\ldots,n$, с которой объект $x_i\in X$, $\forall i=1,\ldots,n$ принадлежит нечеткому кластеру A^l, $l=1,\ldots,c$, так что главной особенностью нечетких методов кластеризации является сочетание высокой точности с содержательной осмысленностью результатов классификации.

Наиболее распространенным подходом к решению нечеткой модификации задачи автоматической классификации является оптимизационный подход, методы которого отыскивают экстремум некоторого критерия качества классификации, примером которого может послужить критерий Дж. Беждека:

$$Q_B(P,\overline{\mathrm{T}})=\sum_{l=1}^{c}\sum_{i=1}^{n}u_{li}^{\gamma}\left\|x_i-\overline{\tau}^l\right\|^2, \qquad (1)$$

где c – число нечетких кластеров в искомом нечетком c-разбиении P; $1<\gamma<\infty$ – показатель, определяющий степень нечеткости классификации; $\overline{\mathrm{T}}=\{\overline{\tau}^1,\ldots,\overline{\tau}^c\}$ – множество прототипов нечетких кластеров A^l, $l=1,\ldots,c$. Локальный минимум критерия (1) отыскивается при ограничении:

$$\sum_{l=1}^{c}u_{li}=1,\quad l=1,\ldots,c;\quad i=1,\ldots,n, \qquad (2)$$

именуемом в специальной литературе условием нечеткого c-разбиения и являющемся общим для всех оптимизационных методов нечеткой кластеризации. Численная процедура, минимизирующая (1), широко известна в специальной литературе под обозначением FCM-алгоритма и является основой семейства других нечетких кластер-процедур.

Разновидностью оптимизационных методов нечеткой кластеризации являются методы возможностной кластеризации [2], специфика которых заключается в том, что структура, образуемая нечеткими кластерами, удовлетворяет условию возможностного разбиения:

$$\sum_{l=1}^{c}\mu_{li}>1,\quad l=1,\ldots,c;\quad i=1,\ldots,n, \qquad (3)$$

являющегося менее жестким, чем условие нечеткого c-разбиения (2), и значения принадлежности μ_{li}, $l = 1, \ldots, c$, $i = 1, \ldots, n$ интерпретируются как степени типичности объекта x_i для нечеткого кластера, а функция принадлежности интерпретируется как функция распределения возможностей. Методы возможностной кластеризации получают все большее распространение как в теоретических исследованиях, так и на практике в силу их устойчивости к наличию в исследуемой совокупности аномальных наблюдений и простоты интерпретации результатов классификации.

В [3] предложен подход к решению нечеткой модификации задачи автоматической классификации, использующей так называемый механизм частичного обучения, сущность которого заключается в том, что относительно некоторого подмножества $X_L = \{x_{L(1)}, \ldots, x_{L(c)}\}$ объектов исследуемой совокупности $X = \{x_1, ..., x_n\}$ имеется априорная информация об их принадлежности классам A^l, $l = 1, \ldots, c$ нечеткого c-разбиения P, которая может быть использована при построении оптимальной классификации. Иными словами, если X_L – множество помеченных объектов, $X_L \subset X$, элементы которого представлены булевыми векторами $s = (s_1, s_2, \ldots, s_n)^T$, где T – символ транспонирования и $s_{li} = 1$, если $x_i \in X_L$ и объект x_i является меткой для нечеткого кластера A^l, $l \in \{1, \ldots, c\}$, т. е. $x_i = x_{L(l)}$; в противном случае, если $x_i \notin X_L$, то имеет место $s_{li} = 0$. В свою очередь $Y_{c \times n} = [y_{li}]$, $l = 1, \ldots, c$; $i = 1, \ldots, n$ – матрица нечеткого c-разбиения, составляемая исследователем в соответствии со следующим правилом: если $x_i \in X_L$, то y_{li} задается исследователем с соблюдением условия $\sum_{l=1}^{c} y_{li} = 1$, где y_{li} – степень принадлежности помеченного объекта x_i, $x_i \in X_L$ классу A^l, $l = 1, \ldots, c$; иначе, при $x_i \notin X_L$ соответствующий столбец в матрице $Y_{c \times n}$ оказывается не нужным и пропускается при обработке матрицы $Y_{c \times n}$. В таком случае задача классификации состоит в минимизации критерия вида

$$Q_P(P, \overline{\mathrm{T}}) = \sum_{l=1}^{c} \sum_{i=1}^{n} u_{li}^2 \left\| x_i - \overline{\tau}^l \right\|^2 + \sum_{l=1}^{c} \sum_{i=1}^{n} (u_{li} - s_{li} y_{li})^2 \left\| x_i - \overline{\tau}^l \right\|^2 \qquad (4)$$

при ограничении (2).

В [3] предложены различные модификации критерия (4), одна из которых базируется на взвешивании в (4) обоих слагаемых, а другая – с заменой в качестве функции расстояния квадрата евклидовой нормы на квадрат расстояния Махаланобиса. С содержательной точки зрения, минимизация первого слагаемого в (4), полностью совпадающего с критерием (1) при $\gamma = 2$, минимизирует нечеткие суммы квадратов расстояний от объектов до прототипов нечетких кластеров, а второе слагаемое в (4) является взвешенной по квадратам расстояний суммой отклонений расчетных значений функции принадлежности объектов нечетким кластерам от заданных априорно. Очевидно, что помеченные объекты частично определяют структуру строящейся классификации исследуемой совокупности X, и множество X_L может интерпретироваться как частично обучающая выборка, элементы которого являются эталонами для классификации. Однако следует указать, что выбор экспертом помеченных объектов и априорных значений принадлежности существенно влияет на результат классификации.

Априорная информация о принадлежности некоторых объектов исследуемой совокупности классам искомого нечеткого c-разбиения позволяет значительно повысить как точность классификации, так и скорость сходимости кластер-процедуры, что также демонстрируется в [3], в силу чего подход к нечеткой кластеризации, использующей аппарат частичного обучения, получил дальнейшее развитие, а соответствующие методы широко внедряются при решении разнообразных задач [4, 5].

Как отмечалось выше, наибольшее распространение получили оптимизационные методы нечеткой кластеризации, вводящие задачу классификации в сугубо математическое русло, однако эвристические методы нечеткой кластеризации, несмотря на меньшее распространение, являются также удобным инструментом анализа данных в силу их простоты и наглядности. В [6] предложен эвристический метод нечеткой кластеризации, заключающийся в построении так называемого распределения по априори задаваемому числу c нечетких α-кластеров, удовлетворяющих введенному определению. В свою очередь в [7] было продемонстрировано, что распределение по нечетким α-кластерам является частным случаем возможностного разбиения (3), и соответствующая процедура, как и ее последующие модификации, представляет собой эвристический алгоритм возможностной кластеризации, в силу чего предложенная в [6] версия алгоритма, от аббревиатуры английских терминов direct – прямой и allotment among fuzzy clusters – распределение по нечетким кластерам, получила обозначение D-AFC(c)-алгоритма. Если $X = \{x_1, ..., x_n\}$ – совокупность объектов, на которой определена нечеткая толерантность T с функцией принадлежности $\mu_T(x_i, x_j)$, $i, j = 1, \ldots, n$, т. е. бинарное нечеткое отношение на X, удовлетворяющее условиям симметричности и рефлексивности, и информация о совокупности X представлена в виде матрицы коэффициентов близости $\rho_{n\times n} = [\mu_T(x_i, x_j)]$, так что строки или столбцы этой матрицы являются нечеткими множествами $\{A^1, ..., A^n\}$, то для некоторого α, $\alpha \in (0,1]$, нечеткое множество уровня α, определяемое условием $A^l_{(\alpha)} = \{(x_i, \mu_{A^l}(x_i)) \mid \mu_{A^l}(x_i) \geq \alpha\}$, $l \in [1, n]$, такое, что $A^l_{(\alpha)} \subseteq A^l$, $A^l \in \{A^1, \ldots, A^n\}$, будет называться нечетким α-кластером с функцией принадлежности μ_{li} объекта $x_i \in X$ нечеткому α-кластеру $A^l_{(\alpha)}$, определяемой выражением

$$\mu_{li} = \begin{cases} \mu_{A^l}(x_i), & x_i \in A^l_\alpha; \\ 0 - \text{в противном случае}, \end{cases} \qquad (5)$$

где $A^l_\alpha = \{x_i \in X \mid \mu_{A^l}(x_i) \geq \alpha\}$ – α-уровень A^l, $l \in \{1, \ldots, n\}$. Объект $x_i \in X$, обладающий наибольшим значением функции принадлежности μ_{li} некоторому нечеткому α-кластеру $A^l_{(\alpha)}$, именуется его типичной точкой и обозначается τ^l, а функция принадлежности, определяемая выражением (5), показывает степень сходства i-го объекта множества X с типичной точкой τ^l соответствующего нечеткого α-кластера. Если условие (3) выполняется для всех $A^l_{(\alpha)} \in R^\alpha_z(X)$, где $R^\alpha_z(X) = \{A^l_{(\alpha)} \mid l = \overline{1, c},\ 2 \leq c \leq n\}$ – семейство c нечетких α-кластеров для некоторого значения α, порожденных заданной на X нечеткой толерантностью T, то это семейство является распределением множества классифицируемых объектов X по c нечетким α-кластерам. Условие (3) в рассматриваемом случае требует, чтобы все объекты совокупности X были распределены по c нечетким α-кластерам $\{A^1_{(\alpha)}, ..., A^c_{(\alpha)}\}$ с положительными значениями μ_{li}, $l = 1, \ldots, c$, $i = 1, \ldots, n$.

Сущность D-AFC(c)-алгоритма заключается в построении множества допустимых решений $B(c) = \{R^\alpha_z(X)\}$ для c классов с последующим выбором в качестве решения задачи классификации некоторого единственного распределения $R^*(X) \in B(c)$. Выбор $R^*(X)$ основывается на вычислении для всех $R^\alpha_z(X) \in B(c)$ критерия

$$F(R^\alpha_z(X), \alpha) = \sum_{l=1}^{c} \frac{1}{n_l} \sum_{i=1}^{n_l} \mu_{li} - \alpha c, \qquad (6)$$

определяющего качество каждого $R^\alpha_z(X) \in B(c)$, где $n_l = card(A^l_\alpha)$ – мощность носителя нечеткого множества $A^l_{(\alpha)} \in R^\alpha_z(X)$, $l \in \{1, \ldots, c\}$, $\alpha \in (0, 1]$, так что (6) определяет среднюю суммарную принадлежность объектов множества X нечетким α-кластерам $\{A^1_{(\alpha)}, ..., A^c_{(\alpha)}\}$ распределения $R^\alpha_z(X)$ за вычетом величины αc, регуляризующей число классов в $R^\alpha_z(X)$, и оптимальному распределению $R^*(X)$ соответствует максимальное значение (6), так что решение состоит в построении распределения, удовлетворяющего условию

$$R^*(X) = \arg \max_{R^\alpha_z(X) \in B(c)} F(R^\alpha_z(X), \alpha). \qquad (7)$$

Результатом работы D-AFC(c)-алгоритма является не только распределение $R^*(X)$ объектов совокупности X по заданному числу c нечетких α-кластеров, но и соответствующее значение порога сходства α.

Как указывалось выше, D-AFC(c)-алгоритм представляет собой базовую версию кластер-процедуры. В работе [7] предлагается его модификация, использующая аппарат частичного обучения, в силу чего (partial supervision – частичное обучение) получившая обозначение D-AFC-PS(c)-алгоритма. Механизм частичного обучения, используемый в D-AFC-PS(c)-алгоритме, достаточно прост: если $X_L = \{x_{L(1)}, \ldots, x_{L(c)}\}$ – множество помеченных объектов, и объект $x_i \in X_L$ является меткой для нечеткого α-кластера $A^l_{(\alpha)}$, $l \in \{1, \ldots, c\}$, т. е. $x_i = x_{L(l)}$, то априорное значение принадлежности y_{li} помеченного объекта x_i соответствующему $A^l_{(\alpha)}$, $l \in \{1, \ldots, c\}$ задается исследователем, при этом $card(X_L) = c$, т. е. общее количество помеченных объектов равно числу c нечетких α-кластеров в искомом распределении $R^*(X)$, и каждый помеченный объект должен быть распределен в единственный нечеткий α-кластер,

а результирующее значение принадлежности μ_{li} помеченного объекта x_i нечеткому α-кластеру $A^l_{(\alpha)}$, $l \in \{1, \ldots, c\}$ должно быть не меньшим, чем заданное априорно y_{li}. По сравнению с методом, используемым в алгоритме В. Педрича, метод частичного обучения, используемый в D-AFC-PS(c)-алгоритме, очевидно, является менее громоздким, простым в реализации и ясным с содержательной точки зрения.

Вместе с тем при решении задач, требующих высокой точности классификации в условиях ограниченного лимита времени, что имеет большое значение в системах поддержки принятия решений специального назначения, помимо экспертного знания о принадлежности объектов классам, используемого при построении множества $X_L = \{x_{L(1)}, \ldots, x_{L(c)}\}$ и задании априорных значений принадлежности y_{li} для элементов X_L, оказывается необходимым проведение предварительного анализа исследуемой совокупности с целью получения обучающей информации для последующего применения методов нечеткой кластеризации с частичным обучением. Указанный подход, основанный на предварительной обработке исследуемой совокупности с помощью D-AFC(c)-алгоритма и выбором в качестве помеченных объектов типичных точек $\{\tau^1, \ldots, \tau^c\}$ нечетких α-кластеров $A^l_{(\alpha)}$, $l = 1, \ldots, c$, полученного распределения $R^*(X)$ с последующей обработкой данных алгоритмом В. Педрича, был предложен в [8] и продемонстрировал высокую эффективность. В [9] предложен подход к построению множества X_L и соответствующих значений y_{li} для использования в D-AFC-PS(c)-алгоритме, основанный на предварительной обработке данных об X некоторой оптимизационной нечеткой кластер-процедурой с последующим вычислением расстояния $d(x_i, \overline{\tau}^l)$ от всех объектов $x_i \in X$ до прототипов $\{\overline{\tau}^1, \ldots, \overline{\tau}^c\}$ кластеров A^l, $l = 1, \ldots, c$ нечеткого c-разбиения P, нормировкой $\tilde{d}(x_i, \overline{\tau}^l) = d(x_i, \overline{\tau}^l) \Big/ \left(\max\limits_i d(x_i, \overline{\tau}^l) \right)$ и вычислением коэффициентов близости $\tilde{s}(x_i, \overline{\tau}^l) = 1 - \tilde{d}(x_i, \overline{\tau}^l)$, так что объекты, находящиеся наиболее близко к прототипам, могут быть выбраны в качестве помеченных, а соответствующие значения $\tilde{s}(x_i, \overline{\tau}^l)$ – в качестве априорных значений принадлежности y_{li}.

Подходы, предложенные в [8, 9], требуют априорного знания о числе c классов в искомом нечетком c-разбиении P или распределении по нечетким α-кластерам $R^*(X)$. В ряде ситуаций оказывается необходимым построить максимально точную классификацию в условиях полного отсутствия информации об исследуемой совокупности X. В таком случае вначале представляется целесообразной обработка X кластер-процедурой, автоматически определяющей число классов c, с последующим выделением множества X_L с соответствующими значениями y_{li}, $l \in \{1, \ldots, c\}$, для чего можно воспользоваться предложенной в [10] модификацией D-AFC(c)-алгоритма, использующей транзитивное замыкание нечеткой толерантности, в силу чего – от аббревиатуры выражения transitive closure – получившей условное обозначение D-AFC-TC-алгоритма. Так как транзитивное замыкание нечеткой толерантности представляет собой нечеткую эквивалентность, разбивающую предметную область на непересекающиеся классы, для распределений $R_z^{\alpha}(X)$ различных уровней α число нечетких кластеров c будет различным, и задачей классификации является выделение априори неизвестного числа нечетких α-кластеров, для чего в последовательности $0<\alpha_0<\ldots<\alpha_l<\ldots<\alpha_Z=1$ на основе вычисления скачка значений порога α определяется такое значение α_l, которому соответствует некоторое неизвестное число нечетких α-кластеров c. Помимо того, что D-AFC-TC-алгоритм отыскивает априори неизвестное число c нечетких α-кластеров, отличающих его от D-AFC(c)-алгоритма, особенностями является, во-первых, то, что для D-AFC-TC-алгоритма матрицей исходных данных является матрица «объект–признак», и для решения задачи классификации используются как критерий (6), так и некоторая метрика $d(x_i, x_j)$, а, во-вторых, то обстоятельство, что результатом работы D-AFC-TC-алгоритма будут также координаты прототипов $\{\overline{\tau}^1, \ldots, \overline{\tau}^c\}$ нечетких α-кластеров $\{A^1_{(\alpha)}, ..., A^c_{(\alpha)}\}$ распределения $R^*(X)$. В силу того что транзитивное замыкание нечеткой толерантности искажает геометрическую структуру исследуемой совокупности X, D-AFC-TC-алгоритм оказывается полезным только на этапе разведочного анализа данных. Таким образом, сущность предлагаемого метода частичного обучения для использования в D-AFC-PS(c)-алгоритме в условиях отсутствия информации о числе классов c, на которые «расслаивается» множество объектов X, заключается в построении с помощью D-AFC-TC-алгоритма распределения $R^*(X)$ по неизвестному числу c нечетких α-кластеров с последующим выбором в качестве элементов множества X_L типичных точек $\{\tau^1, \ldots, \tau^c\}$ нечетких α-кластеров. В качестве значения y_{li}, $l\in\{1, \ldots, c\}$, общего для всех помеченных объектов, целесообразно выбрать полученное в результате работы D-AFC-TC-алгоритма значение порога сходства α, так как при обработке данных D-AFC-PS(c)-алгоритмом геометрическая структура X не претерпевает изменений, и типичными точками классов распределения $R^*(X)$, полученного с помощью D-AFC-PS(c)-алгоритма, могут оказаться другие объекты.

Эффективность предложенного подхода к построению подмножества помеченных объектов и определению априори задаваемой функции принадлежности для использования в D-AFC-PS(c)-алгоритме целесообразно проиллюстрировать на простом примере. Для проведения вычислительного эксперимента были выбраны представленные на рис. 1 двумерные данные о 15 объектах, предложенные в [11].

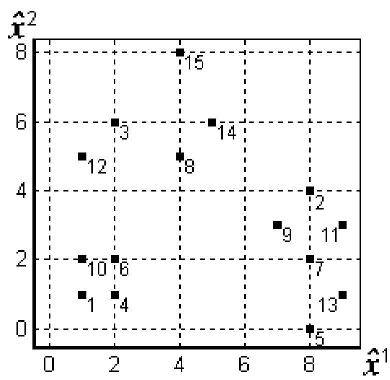

Рис. 1. Двумерные данные для проведения вычислительного эксперимента

На рис. 1 визуально выделяются три группы объектов – $\{x_1, x_4, x_6, x_{10}\}$; $\{x_3, x_8, x_{12}, x_{14}, x_{15}\}$ и $\{x_2, x_5, x_7, x_9, x_{11}, x_{13}\}$, которые в дальнейшем будут использованы для верификации результатов вычислительных экспериментов. Обозначая объекты символами x_i, $i=1, \ldots, 15$, а признаки – символами $\hat{x}^t$, $t=1, 2$, была получена матрица «объект–признак» $X_{15\times 2}=[\hat{x}_i^t]$, которая обработана с помощью нормализации [12]:

$$x_i^t=\frac{\hat{x}_i^t}{\max\limits_i \hat{x}_i^t},\quad i=1, \ldots, n;\quad t=1, \ldots, m, \qquad (8)$$

вследствие чего каждый объект может интерпретироваться как нечеткое множество на универсуме признаков с функцией принадлежности $\mu_{x_i}(x^t)$, $i = 1, \ldots, n$, с последующим применением квадрата относительного евклидова расстояния между нечеткими множествами [10]

$$e^2(x_i, x_j) = \frac{1}{m}\sum_{t=1}^{m}\left(\mu_{x_i}(x^t) - \mu_{x_j}(x^t)\right)^2,$$
$$i, j = 1, \ldots, n; \quad t = 1, \ldots, m, \tag{9}$$

и операции дополнения $\mu_T(x_i, x_j) = 1 - e^2(x_i, x_j)$, $i, j = 1, \ldots, 15$, была построена матрица нечеткой толерантности $T_{15\times 15} = [\mu_T(x_i, x_j)]$, результатом обработки которой с помощью D-AFC(c)-алгоритма при числе классов $c = 3$ является распределение $R^*(X)$ по полностью разделенным нечетким α-кластерам, полученное при значении порога сходства $\alpha = 0,7912$. Значения принадлежности объектов исследуемой совокупности нечетким α-кластерам представлены на рис. 2.

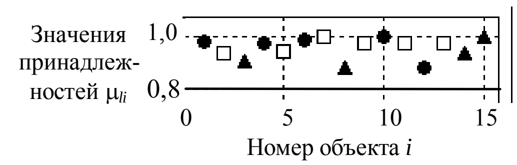

Рис. 2. Результат обработки множества объектов D-AFC(c)-алгоритмом

На рис. 2 и последующих рисунках значения принадлежностей объектов 1-му классу обозначены символом «●», 2-му – символом «▲», и 3-му – символом «□». Анализ представленного на рис. 2 результата классификации позволяет выделить в качестве типичной точки τ^1 первого класса объект x_{10}, типичной точки τ^2 второго – объект x_{15}, а для третьего класса имеет место $\tau^3 = x_7$; в свою очередь носители нечетких α-кластеров полученного распределения $R^*(X)$ образуют группы $\{x_1, x_4, x_6, x_{10}, x_{12}\}$, $\{x_3, x_8, x_{14}, x_{15}\}$ и $\{x_2, x_5, x_7, x_9, x_{11}, x_{13}\}$, что ввиду отнесения объекта x_{12} к 1-му классу не совпадает с визуальным выделением классов на рис. 1.

В результате обработки исходных данных D-AFC-TC-алгоритмом с помощью нормировки (8) и расстояния (9) было получено распределение $R^*(X)$ также по трем нечетким α-кластерам при значении порога сходства $\alpha =$ $= 0,9609$, значения принадлежности объектов которым изображены на рис. 3.

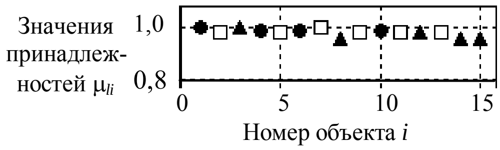

Рис. 3. Результат обработки множества объектов D-AFC-TC-алгоритмом

Носители нечетких α-кластеров представляют собой подмножества $\{x_1, x_4, x_6, x_{10}\}$, $\{x_3, x_8, x_{12}, x_{14}, x_{15}\}$ и $\{x_2, x_5, x_7, x_9, x_{11}, x_{13}\}$, соответствующие визуально выделенным на рис. 1 классам, а типичными точками нечетких α-кластеров являются объекты $\tau^1 = x_1$, $\tau^2 = x_3$ и $\tau^3 = x_7$ соответственно. Таким образом, соответствующие объекты были выбраны в качестве помеченных с общим для всех значением априорной функции принадлежности $y_{li} =$ $= 0,9609$, $l = 1, \ldots, 3$, $i = 1, \ldots, 3$, для обработки тестовых данных с помощью D-AFC-PS(c)-алгоритма. Значения принадлежностей объектов нечетким α-кластерам распределения $R^*(X)$, построенного с помощью D-AFC-PS(c)-алгоритма, изображены на рис. 4.

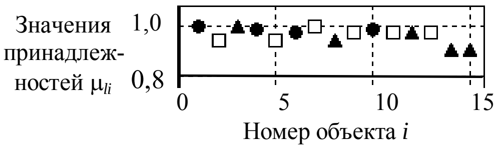

Рис. 4. Результат обработки множества объектов D-AFC-PS(c)-алгоритмом

Значение порога сходства при обработке данных с помощью D-AFC-PS(c)-алгоритма составило $\alpha = 0,8220$, а выделение носителей нечетких α-кластеров дает классы $\{x_1, x_4, x_6, x_{10}\}$, $\{x_3, x_8, x_{12}, x_{14}, x_{15}\}$ и $\{x_2, x_5, x_7, x_9, x_{11}, x_{13}\}$, соответствующие визуально выделенным классам. Кроме того, в этом эксперименте, как и при обработке данных D-AFC-TC-алгоритмом, типичными точками нечетких α-кластеров являются объекты $\tau^1 = x_1$, $\tau^2 = x_3$ и $\tau^3 = x_7$, которые наименее удалены от геометрических центров соответствующих групп. Таким образом, вычислительный эксперимент наглядно демонстрирует не только преимущество использования механизма частичного обучения при обращении к эвристическому методу нечеткой кластеризации для решения задач классификации, но и эффективность предложенного метода частичного обучения.

Анализ результатов, полученных с помощью D-AFC(c)-алгоритма и D-AFC-PS(c)-алгоритма, проводился в сравнении с оптимизационными алгоритмами нечеткой кластеризации – FCM-алгоритмом и алгоритмом В. Педрича [3], минимизирующим критерий (4), при этом в обоих экспериментах полагалось $c = 3$, а в эксперименте с FCM-алгоритмом значение показателя нечеткости γ полагалось равным двум. Значения принадлежностей объектов нечетким кластерам, полученным с помощью FCM-алгоритма, изображены на рис. 5.

Интерпретация результатов классификации с помощью правила наибольшей принадлежности приводит к выделению групп $\{x_1, x_4, x_6, x_{10}\}$, $\{x_3, x_8, x_{12}, x_{14}, x_{15}\}$ и $\{x_2, x_5, x_7, x_9, x_{11}, x_{13}\}$, что совпадает с визуально выделенными на рис. 1 классами и результатами обработки данных D-AFC-PS(c)-алгоритмом. Однако следует отметить сравнительно невысокое значение принадлежности объекта x_{12} второму нечеткому кластеру.

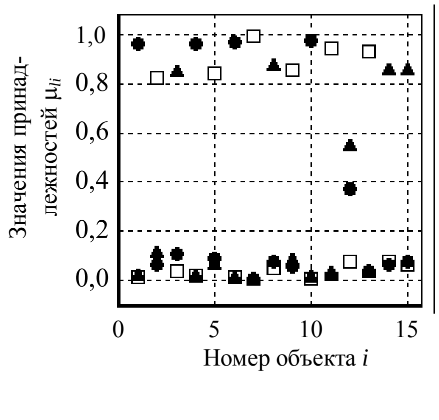

Рис. 5. Результат обработки множества объектов FCM-алгоритмом

Обработка данных алгоритмом В. Педрича проводилась с помощью обучающей информации, использовавшейся при их обработке D-AFC-PS(c)-алгоритмом. Но так как обращение к алгоритму В. Педрича подразумевает использование в качестве обучающей информации матрицы нечеткого c-разбиения $Y_{c\times n} = [y_{li}]$, для ее построения значения y_{li} принадлежностей помеченного объекта классам, для которых он не является меткой, вычислялись по формуле $y_{li} = (1-\alpha)/(c-1)$, что обеспечивает выполнение условия нечеткого c-разбиения для $Y_{c\times n}$. Значения принадлежностей объектов классам нечеткого c-разбиения $P_{c\times n} = [u_{li}]$, полученного при обработке тестовых данных алгоритмом В. Педрича, изображены на рис. 6.

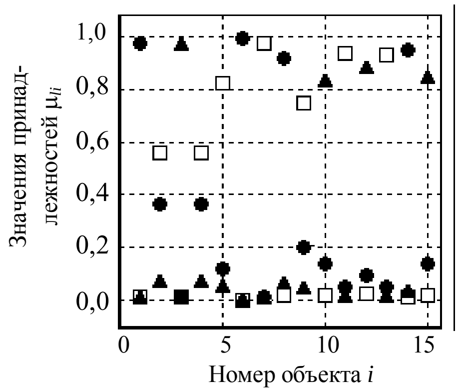

Рис. 6. Результат обработки множества объектов алгоритмом В. Педрича

Как и в случае эксперимента с FCM-алгоритмом, результат классификации интерпретировался на основе правила наибольшей принадлежности, что позволило выделить группы $\{x_1, x_6, x_8, x_{14}\}$, $\{x_3, x_{10}, x_{12}, x_{15}\}$ и $\{x_2, x_4, x_5, x_7, x_9, x_{11}, x_{13}\}$. Подобное искажение результатов классификации в сравнении с FCM-алгоритмом объясняется выбором нормализации (8), достаточно сильно искажающей геометрию исходных данных, для нормировки исходных данных при их обработке алгоритмом В. Педрича – на это обстоятельство указывают и одинаковые значения принадлежностей объектов x_2 и x_4 всем трем классам полученного нечеткого c-разбиения. В свою очередь, использование унитаризации [12] для нормировки данных при сохранении прежней обучающей информации приводит к результатам, сходным с результатами обработки исходных данных FCM-алгоритмом, что свидетельствует о высокой чувствительности алгоритма В. Педрича к выбору способа нормировки. Кроме того, очевидно, что использованный способ задания априорных значений принадлежности для помеченных объектов в алгоритме В. Педрича недостаточно адекватен в силу различия условий нечеткого c-разбиения (2) и возможностного разбиения (3).

ВЫВОД

В работе предложен метод построения подмножества помеченных объектов и соответствующих априорных значений принадлежности для использования в эвристическом алгоритме возможностной кластеризации с частичным обучением, основой которого является предварительная обработка данных с помощью модификации эвристического алгоритма возможностной кластеризации, не требующей задания параметров, что делает предложенный метод пригодным в условиях полного отсутствия априорной информации о структуре исследуемой совокупности. Анализ результатов вычислительных экспериментов наглядно демонстрирует высокую эффективность метода, использующего аппарат частичного обучения, в сравнении с базовой версией метода, а также нечеткими кластер-процедурами. Следует также отметить, что предложенная схема двухэтапной возможностной кластеризации позволяет производить классификацию данных в полностью автоматическом режиме.

ЛИТЕРАТУРА

1. **Bezdek, J. C.** Pattern recognition with fuzzy objective function algorithms / J. C. Bezdek. – New York: Plenum Press, 1981. – 230 p.

2. **Krishnapuram, R.** A possibilistic approach to clustering / R. Krishnapuram, J. M. Keller // IEEE Transactions on Fuzzy Systems. – 1993. – Vol. 1. – P. 98–110.

3. **Pedrycz, W.** Algorithms of fuzzy clustering with partial supervision / W. Pedrycz // Pattern Recognition Letters. – 1985. – Vol. 3. – P. 13–20.

4. **Abonyi, J.** Supervised fuzzy clustering for the identification of fuzzy classifiers / J. Abonyi, F. Szeifert // Pattern Recognition Letters. – 2003. – Vol. 24. – P. 2195–2207.

5. **Liu, H.** Evolutionary semi-supervised fuzzy clustering / H. Liu, S.T. Huang // Pattern Recognition Letters. – 2003. – Vol. 24. – P. 3105–3113.

6. **Viattchenin, D. A.** A new heuristic algorithm of fuzzy clustering / D. A. Viattchenin // Control & Cybernetics. – 2004. – Vol. 33. – P. 323–340.

7. **Viattchenin, D. A.** A direct algorithm of possibilistic clustering with partial supervision / D. A. Viattchenin // Journal of Automation, Mobile Robotics and Intelligent Systems. – 2007. – Vol. 1. – P. 29–38.

8. **Viattchenin, D. A.** A methodology of fuzzy clustering with partial supervision / D. A. Viattchenin // Systems Science. – 2007. – Vol. 33. – P. 61–71.

9. **Viattchenin, D. A.** Fuzzy objective function-based technique of partial supervision for a heuristic method of possibilistic clustering / D. A. Viattchenin // Neural Networks and Artificial Intelligence: Proceedings of the Fifth International Conference ICNNAI'2008. – Minsk, 2008. – P. 51–55.

10. **Вятченин, Д. А.** Прямые алгоритмы нечеткой кластеризации, основанные на операции транзитивного замыкания и их применение к обнаружению аномальных наблюдений / Д. А. Вятченин // Искусственный интеллект. – 2007. – № 3. – С. 205–216.

11. **Looney, C. G.** Interactive clustering and merging with a new fuzzy expected value / C. G. Looney // Pattern Recognition. – 2002. – Vol. 35. – P. 2413–2423.

12. **Walesiak, M.** Ugólniona miara odległości w statystycznej analizie wielowymiarowej / M. Walesiak. – Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, 2002. – 107 s.

Поступила 23.03.2009