УДК 510.22+519.237.8

ПОСТРОЕНИЕ РАСПРЕДЕЛЕНИЯ ПО НЕЧЕТКИМ КЛАСТЕРАМ В СЛУЧАЕ КВАЗИУСТОЙЧИВОЙ КЛАСТЕРНОЙ СТРУКТУРЫ МНОЖЕСТВА ОБЪЕКТОВ

Д.А. ВЯТЧЕНИН, А.В. ДОМОРАЦКИЙ

Объединенный институт проблем информатики НАН Беларуси
Сурганова 6, 220012, Минск, Беларусь

НИРУП "Геоинформационные системы" НАН Беларуси
Сурганова 6, 220012, Минск, Беларусь

Поступила в редакцию 19 октября 2009

Предложен метод кластеризации объектов с варьирующимися в интервале значениями признаков в случаях устойчивой или квазиустойчивой кластерной структуры множества объектов.

Ключевые слова: динамические признаки, интервально-значное нечеткое множество, возможностная кластеризация, распределение по нечетким кластерам, типичная точка.

Введение

При решении задач автоматической классификации динамических объектов, т.е. объектов, признаки которых могут изменять свои значения с течением времени или при наличии внешних воздействий [1], традиционно используются различные подходы, основанные на методах нечеткой и возможностной кластеризации [2], в которых 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 $\mu_{li} \in [0,1]$, $l = 1, \ldots, c$, $i = 1, \ldots, n$ с которой объект $x_i \in X$ $\forall i = 1, \ldots, n$, принадлежит тому или иному нечеткому кластеру $A^l, l = 1, \ldots, c$.

В задачах динамической кластеризации признаки $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$ объектов $x_i \in X$ могут принимать значения в непрерывном интервале безотносительно к моменту измерения соответствующей характеристики объекта, так что каждый признак $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$, для объекта x_i, $i = 1, \ldots, n$ представляет собой интервал значений $[\tilde{x}_i^{t_1 \min}, \tilde{x}_i^{t_1 \max}]$. Кластерная структура исследуемой совокупности, состоящей из подобных объектов, также является динамической, и зависит от значений признаков в момент классификации. На содержательном уровне задача построения устойчивой кластерной структуры в [1] формулируется следующим образом: найти такое априори неизвестное число c областей признакового пространства $\Re^m$, в которых отображаются кластеры, при различных значениях принимаемых объектами исследуемой совокупности X признаков $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$, варьирующихся в интервале $[\tilde{x}_i^{t_1 \min}, \tilde{x}_i^{t_1 \max}]$. В свою очередь, перед решением указанной задачи сначала необходимо установить тип динамических изменений кластерной структуры, для чего в [1] определены понятия устойчивой, квазиустойчивой и неустойчивой кластерной структуры. Если при изменении в соответствующем интервале $[\tilde{x}_i^{t_1 \min}, \tilde{x}_i^{t_1 \max}]$ значений признаков $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$ объектов $x_i \in X$ исследуемой совокупности число c кластеров $\{A^1,...,A^c\}$ не изменяется, и не изменяются координаты их прототипов $\{\overline{\tau}^1,\ldots,\overline{\tau}^c\}$, то структура, образуемая кластерами $\{A^1,...,A^c\}$, называется устойчивой; если с изменением значений признаков объектов число c кластеров $\{A^1,...,A^c\}$ не изменяется, но изменяются координаты их прототипов $\{\overline{\tau}^1,\ldots,\overline{\tau}^c\}$, то соответствующая кластерная структура именуется квазиустойчивой, а если при изменении значений признаков наблюдаемых объектов $x_i \in X$ изменяется число c кластеров, то кластерная структура является неустойчивой. В [1] представлен метод определения типа кластерной структуры совокупности объектов с варьирующимися в интервале значениями признаков, в основе которого лежит D-AFC-TC-алгоритм [3] построения распределения объектов по априори неизвестному числу нечетких α-кластеров.

В случае, когда кластерная структура, образуемая объектами исследуемой совокупности, является неустойчивой, ей соответствуют такие типы динамических изменений, как образование новых кластеров, слияние кластеров, их расщепление и элиминация, а в случае квазиустойчивости кластерной структуры число кластеров не изменяется, однако имеет место дрейф прототипов кластеров, и, как продемонстрировано в [1], изменяются типичные точки кластеров. В отличие от ситуации неустойчивой кластерной структуры, где ее изменения носят скачкообразный характер, в ситуации квазиустойчивой кластерной структуры изменения носят непрерывный, и, как следствие, латентный характер. Указанное обстоятельство позволяет выделить задачу построения распределения объектов по классам в случае квазиустойчивости кластерной структуры в качестве первоочередной.

В настоящем исследовании изложен метод построения распределения объектов, описываемых динамическими признаками, по нечетким α-кластерам в случае, когда кластерная структура является устойчивой или квазиустойчивой. Основой предлагаемого метода является представление объектов исследуемой совокупности как интервально-значных нечетких множеств с последующим построением матрицы нечеткой толерантности на соответствующем универсуме, и обработкой полученных таким образом данных с помощью D-AFC(c)-алгоритма возможностной кластеризации [3].

Метод предварительной обработки интервально-значных данных

Эвристический метод возможностной кластеризации, предложенный в [4], основные понятия которого рассмотрены также в [1], заключается в построении так называемого распределения $R^*(X)$ по априори задаваемому числу c нечетких кластеров. Базовая версия кластер-процедуры, получившая в специальной литературе обозначение D-AFC(c)-алгоритма, требует, чтобы исходные данные об исследуемой совокупности объектов $X = \{x_1,...,x_n\}$ были представлены в виде матрицы $T_{n\times n} = [\mu_T(x_i, x_j)]$ нечеткого отношения толерантности, то есть нечеткого отношения, удовлетворяющего свойствам симметричности и рефлексивности, определенного на соответствующем универсуме. Иными словами, матрица $T_{n\times n} = [\mu_T(x_i, x_j)]$ представляет собой матрицу попарной близости объектов, соответствующие элементы которой принимают значения в интервале $[0,1]$. В случае, когда кластерная структура исследуемой совокупности, признаки объектов которой принимают значения в интервале, является устойчивой или квазиустойчивой, число классов c в искомом распределении $R^*(X)$ может быть установлено с помощью предложенного в [1] метода. Задача, таким образом, заключается в построении на множестве $X = \{x_1,...,x_n\}$ динамических объектов нечеткого отношения толерантности T для последующей обработки полученной матрицы D-AFC(c)-алгоритмом с числом классов c, установленным на этапе анализа устойчивости кластерной структуры [1]. С этой целью представляется целесообразным прибегнуть к аппарату так называемых интервально-значных нечетких множеств [5].

Если X — некоторый универсум, то нечеткое множество A определенное на X, чьи значения функции принадлежности представляют собой фиксированные интервалы из отрезка

$[0,1]$, так что функция принадлежности A, задается отображением $\mu_A : X \to 2^{[0,1]}$, то A именуется нечетким множеством с функцией принадлежности, принимающей значения в интервале, и для обозначения нечетких множеств подобного типа в зарубежной литературе используется термин interval-valued fuzzy sets [5]. Определенное на универсуме X нечеткое множество A с функцией принадлежности, принимающей значения в интервале, задается двумя функциями принадлежности: $\underline{\mu}_A(x_i)$, определяющей нижнее значение интервала значений принадлежности $x_i \in X$, и $\overline{\mu}_A(x_i)$, задающей верхнее значение, так что $0 \le \underline{\mu}_A(x_i) \le \overline{\mu}_A(x_i) \le 1$, и интервально-значное нечеткое множество A определяется как $A = \{x_i, \mu_A(x_i) = [\underline{\mu}_A(x_i), \overline{\mu}_A(x_i)] \mid x_i \in X, \underline{\mu}_A(x_i), \overline{\mu}_A(x_i) \in [0,1]\}$. Очевидно, что каждое обычное нечеткое множество A может быть представлено в виде интервально-значного нечеткого множества с совпадающими для каждого элемента $x_i \in X$ нижним и верхним значениями интервала значений принадлежности, т.е. $\underline{\mu}_A(x_i) = \overline{\mu}_A(x_i)$ $\forall x_i \in X$.

Обозначая объекты исследуемой совокупности символами x_i, $i = 1, \ldots, n$, а признаки — соответственно, символами $\hat{x}^{t_1}$, $t_1 = 1, \ldots, m_1$, матрица "объект-признак" $X_{n \times m_1} = [\hat{x}_i^{t_1 \min} (t_{1\max})]$, где $\hat{x}_i^{t_1 \min} (t_{1\max}) = [\hat{x}_i^{t_1 \min}, \hat{x}_i^{t_1 \max}]$, может быть обработана с помощью обобщенной нормализации

$$x_i^{t_{1\min}(t_{1\max})} = \frac{\hat{x}_i^{t_{1\min}(t_{1\max})}}{\max\limits_{i, t_{1\max}} \hat{x}_i^{t_{1\min}(t_{1\max})}} \tag{1}$$

или обобщенной унитаризации

$$x_i^{t_{1\min}(t_{1\max})} = \frac{\hat{x}_i^{t_{1\min}(t_{1\max})} - \min\limits_{i, t_{1\max}} \hat{x}_i^{t_{1\min}(t_{1\max})}}{\max\limits_{i, t_{1\max}} \hat{x}_i^{t_{1\min}(t_{1\max})} - \min\limits_{i, t_{1\max}} \hat{x}_i^{t_{1\min}(t_{1\max})}}, \tag{2}$$

где $i = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$, предложенных в [6], вследствие чего каждый объект x_i может интерпретироваться как интервально-значное нечеткое множество на универсуме признаков с функцией принадлежности $\mu_{x_i}(x^{t_1}) = [\underline{\mu}_{x_i}(x^{t_1}), \overline{\mu}_{x_i}(x^{t_1})]$, $i = 1, \ldots, n$, где $\underline{\mu}_{x_i}(x^{t_1}) = \mu_{x_i}(x^{t_{1\min}})$ и $\overline{\mu}_{x_i}(x^{t_1}) = \mu_{x_i}(x^{t_1 \max})$.

Для интервально-значных нечетких множеств Х. Юу и Х. Юаном в [7] был определен ряд мер близости. В рассматриваемом случае при представлении объектов исследуемой совокупности $x_i, x_j \in X$ как интервально-значных нечетких множеств x_i и x_j, $i, j = 1, \ldots, n$, определенных на универсуме признаков, меры сходства, введенные в [7], примут вид

$$s_{JY(IVFS)(1)}(x_i, x_j) = 1 - \frac{1}{\sqrt[\lambda]{m_1}} \sqrt[\lambda]{\sum_{t_1=1}^{m_1} \left| \frac{\underline{\mu}_{x_i}(x^{t_1}) + \overline{\mu}_{x_i}(x^{t_1})}{2} - \frac{\underline{\mu}_{x_j}(x^{t_1}) + \overline{\mu}_{x_j}(x^{t_1})}{2} \right|^{\lambda}} \tag{3}$$

и

$$s_{JY(IVFS)(2)}(x_i, x_j) = 1 - \frac{1}{\sqrt[\lambda]{m_1}} \sqrt[\lambda]{\sum_{t_1=1}^{m_1} \left(\frac{\left|\underline{\mu}_{x_i}(x^{t_1}) - \underline{\mu}_{x_j}(x^{t_1})\right|}{2} + \frac{\left|\overline{\mu}_{x_i}(x^{t_1}) - \overline{\mu}_{x_j}(x^{t_1})\right|}{2} \right)^{\lambda}}, \tag{4}$$

где $i, j = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$ и λ — параметр, такой, что $1 \le \lambda < \infty$. Таким образом, значения коэффициентов близости $s_{JY(IVFS)(1)}(x_i, x_j)$ или $s_{JY(IVFS)(2)}(x_i, x_j)$, полученные с помощью выражений (3) или (4) соответственно, будут представлять собой элементы матрицы нечеткой толерантности $T_{n\times n}=[\mu_T(x_i,x_j)]$, являющейся, как указывалось выше, матрицей исходных данных для D-AFC(c)-алгоритма.

В свою очередь, учитывая, что интервально-значные нечеткие множества представляют собой частный случай нечетких множеств типа 2 [8], для построения матрицы исходных данных оказывается возможным применение к нормализованным интервально-значным данным обобщений расстояний для нечетких множеств типа 2, предложенным в [6]. В частности, обобщение нормализованного евклидова расстояния между нечеткими множествами типа 2 x_i и x_j, предложенного в [6], в рассматриваемом случае примет вид

$$e_{G_2}(x_i,x_j)=\sqrt{\frac{1}{m_1}\sum_{t_1=1}^{m_1}\left(\frac{1}{m_2^2}\sum_{u_1,v_1=1}^{m_2=2}\left(\mu_{x_i}(x^{t_1,u_1})-\mu_{x_j}(x^{t_1,v_1})\right)^2\right)}, \quad (5)$$

где индексы u_1 и v_1 используются для обозначения $\underline{\mu}_{x_i}(x^{t_1})$ и $\overline{\mu}_{x_i}(x^{t_1})$ соответственно, и применение которого к интервально-значным нечетким множествам x_i, $i=1,\ldots,n$, позволяет построить матрицу нечеткого отношения несходства $I_{n\times n}=[\mu_I(x_i,x_j)]$. В свою очередь, операция дополнения

$$\mu_T(x_i,x_j)=1-\mu_I(x_i,x_j),\ \forall x_i,x_j,\ i,j=1,\ldots,n, \quad (6)$$

примененная к $I_{n\times n}=[\mu_I(x_i,x_j)]$, дает в результате матрицу слабой нечеткой толерантности $T_{n\times n}=[\mu_T(x_i,x_j)]$, также являющуюся матрицей исходных данных для D-AFC(c)-алгоритма.

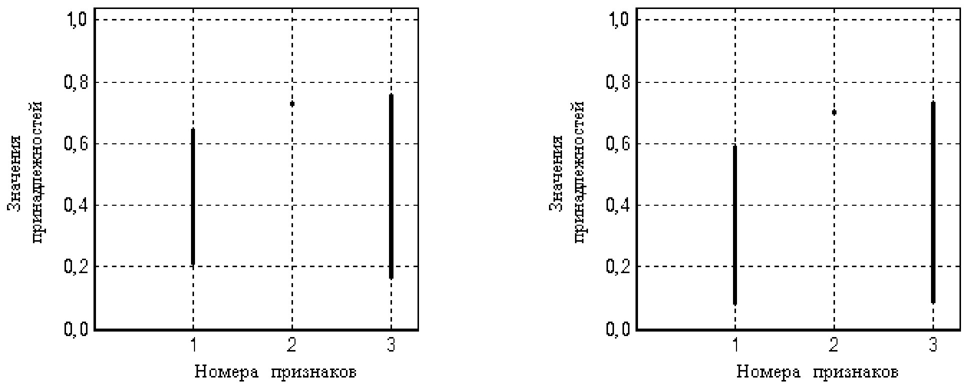
Экспериментальная часть

Для иллюстрации предложенного подхода к построению распределения $R^*(X)$ по заданному числу c нечетких кластеров, целесообразно прибегнуть к тестовым данным М. Сато-Илик и Л. Джейна [9], приведенным в работе [1], где также было определено, что исследуемая совокупность 8 объектов образует квазиустойчивую кластерную структуру с числом классов, равным двум.

Так как различные виды нормировок приводят к различным результатам, вычислительный эксперимент проводился с использованием обоих видов нормировки. К примеру, функции принадлежности $\mu_{x_8}(x^{t_1})=[\underline{\mu}_{x_8}(x^{t_1}),\overline{\mu}_{x_8}(x^{t_1})]$, $t_1=1,\ldots,3$, интервально-значных нечетких множеств, соответствующих восьмому объекту исследуемой совокупности, построенные при использовании нормировок (1) и (2), изображены на рис. 1.

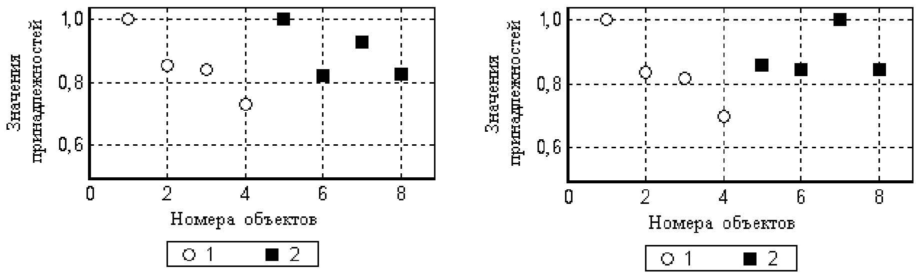
Для построения матрицы нечеткой толерантности $T_{8\times 8}=[\mu_T(x_i,x_j)]$ была выбрана мера сходства (3) при $\lambda=2$. Значения принадлежностей объектов нечетким кластерам распределений $R^*(X)$, полученных в результате обработки матриц $T_{8\times 8}=[\mu_T(x_i,x_j)]$, построенных с помощью нормировок (1) и (2), представлены на рис. 2.

Приведенные на рис. 2 результаты демонстрируют, что при использовании различных видов нормировки изменяются не только значения принадлежности элементов, но и типичная точка τ^2 второго нечеткого кластера, тогда как принадлежности элементов первого класса, значения признаков которых задаются не интервалами, а единичными значениями [9], [1], не претерпевают существенных изменений.



а *б*

Рис. 1. Функции принадлежности интервально-значных нечетких множеств, соответствующих данным о восьмом объекте исследуемой совокупности: *а* — полученные с использованием обобщенной нормализации; *б* — обобщенной унитаризации

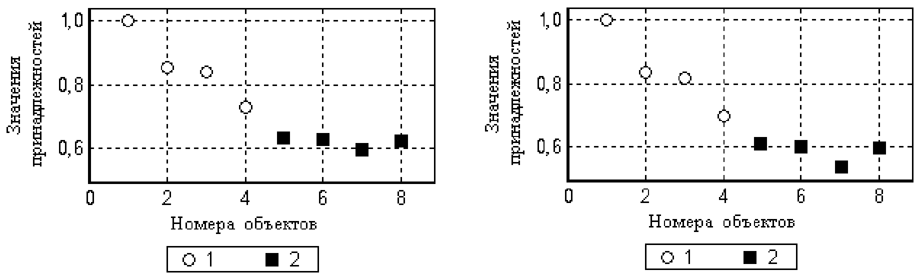


а *б*

Рис. 2. Значения принадлежностей объектов двум классам при кластеризации с использованием меры близости (3) с помощью обобщенной нормализации (*а*) и обобщенной унитаризации (*б*)

Так как результатом работы D-AFC(c)-алгоритма является не только распределение $R^*(X)$ объектов исследуемой совокупности X по заданному числу c нечетких α-кластеров, но и соответствующее значение порога сходства α, необходимо указать, что при использовании для нормировки исходных данных обобщенной нормализации (1) распределение $R^*(X)$ было получено при α=0,5751, а при использовании обобщенной унитаризации (2) значение порога сходства составило α=0,5199. В дополнение к представленным выше результатам для меры сходства (3) следует указать, что при использовании нормировки (1) и меры сходства (4) результаты кластеризации оказались сходными с результатами, полученными при использовании нормировки (2) и меры сходства (3), — так, типичными точками τ^1 и τ^2 нечетких кластеров оказались объекты x_1 и x_7 соответственно, а распределение $R^*(X)$ по двум нечетким кластерам было получено при α=0,5742.

В свою очередь необходимо отметить, что при использовании функции расстояния (5) в сочетании с дополнением (6) для построения матрицы нечеткой толерантности $T_{8\times8}=[\mu_T(x_i,x_j)]$, при использовании как нормировки (1), так и нормировки (2), типичными точками τ^1 и τ^2 нечетких кластеров полученных распределений $R^*(X)$ в обоих случаях оказались объекты x_1 и x_5 — так, при использовании нормировки (1) распределение $R^*(X)$ по двум нечетким кластерам было получено при α=0,5265, а применение нормировки (2) дает в результате распределение $R^*(X)$ по двум нечетким кластерам при значении порога сходства α=0,4933. Следует также указать, что при использовании функции расстояния (5) к матрице нормированных исходных данных вместе с операцией дополнения (6) матрица исходных данных $T_{8\times8}=[\mu_T(x_i,x_j)]$ для D-AFC(c)-алгоритма оказалась матрицей нормальной строгой слабой нечеткой толерантности T_{0n} [10], в силу чего в обоих случаях второй нечеткий кластер оказался слабым нечетким кластером с центром [10], а значения принадлежностей τ^2 составили $\mu_{25} = 0,6404$ и $\mu_{25} = 0,6047$ соответственно. Значения принадлежностей объектов нечетким кластерам распределений $R^*(X)$, полученных в результате обработки матриц $T_{8\times 8} = [\mu_T(x_i, x_j)]$, построенных с помощью нормировок (1) и (2), а также функции расстояния (5) и операции дополнения (6), представлены на рис. 3.



а *б*

Рис. 3. Значения принадлежностей объектов двум классам при кластеризации с использованием формулы (5) при помощи обобщенной нормализации (*а*) и обобщенной унитаризации (*б*)

Следует отметить полное совпадение значений принадлежности объектов первому классу, как в случае эксперимента, результаты которого представлены на рис. 2, так и в случае эксперимента, представленного рис. 3, соответственно.

Для демонстрации преимущества предложенного метода целесообразно привести результаты, полученные М. Сато-Илик и Л. Джейном [8], где исходные данные были представлены в виде матриц предельных значений их признаков, $X^{\min}_{n\times m_1} = [x_i^{t_1 \min}]$ и $X^{\max}_{n\times m_1} = [x_i^{t_1 \max}]$, где $n = 8$ и $m_1 = 3$, после чего была построена матрица $\tilde{X}$ размерности $2n \times m_1$ в соответствии с выражением

$$\tilde{X} \equiv \begin{pmatrix} X^{\min}_{n\times m_1} \\ X^{\max}_{n\times m_1} \end{pmatrix},$$

впоследствии обработанная FANNY-алгоритмом нечеткой кластеризации для числа классов c=2. Значения принадлежностей объектов классам для наименьших и для наибольших значений признаков приведены в таблице.

Результаты обработки тестовых данных FANNY-алгоритмом

Номер объекта	Значения принадлежностей объектов классам			
	для наименьших значений признаков		для наибольших значений признаков	
	1	2	1	2
1	0,80	0,20	0,80	0,20
2	0,76	0,24	0,76	0,24
3	0,83	0,17	0,83	0,17
4	0,77	0,23	0,77	0,23
5	0,23	0,77	0,33	0,67
6	0,29	0,71	0,34	0,66
7	0,27	0,73	0,37	0,63
8	0,27	0,73	0,35	0,65

Анализ результатов, представленных в таблице, показывает, что в методе обработки интервально-значных данных, предложенном в [8], принадлежности динамических объектов классам представляют собой пары значений, $u_{li}^{\min}$ и $u_{li}^{\max}$, $l = 1,\ldots,c$, $i = 1,\ldots,n$, так что результат классификации представляет собой матрицу размерности $2c \times n$, строящуюся в виде

$$\tilde{P} \equiv \begin{pmatrix} P_{c\times n}^{\min} \\ P_{c\times n}^{\max} \end{pmatrix},$$

что затрудняет содержательный анализ результатов классификации. Следует вместе с тем отметить, что для объектов x_i, $i = 1,\ldots,4$, имеет место $u_{li}^{\min} = u_{li}^{\max}$, $l = 1,2$.

Заключение

Анализ приведенных результатов наглядно демонстрирует, что значения типичности μ_{li} в матрице распределения динамических объектов по c классам $R^*(X) = [\mu_{li}]$ размерности $c \times n$ представляют собой единственное значение, что, по сравнению с методом, предложенным М. Сато-Илик и Л. Джейном [8], является более удобным при интерпретации результатов классификации. Если в результате анализа устойчивости кластерной структуры, проведенного с помощью предложенного в [1] подхода, окажется, что кластерная структура исследуемой совокупности является неустойчивой, то для построения распределения $R^*(X)$ по неизвестному числу c нечетких кластеров с помощью предложенного метода вначале необходимо построить множество значений возможного числа классов $c \in \{c_*,\ldots,c^*\}$, где c_* — наименее возможное, а c^* — наиболее возможное число классов в искомом распределении $R^*(X)$, после чего матрица нечеткой толерантности должна быть обработана D-AFC(c)-алгоритмом для всех $c \in \{c_*,\ldots,c^*\}$ с определением оптимального числа c на основе вычисления показателя валидности числа нечетких кластеров.

В работе [11] рассмотрено применение изложенного подхода к решению задачи декомпозиции элементов сложной системы в процессе имитационного моделирования.

CONSTRUCTING OF ALLOTMENT AMONG FUZZY CLUSTERS IN CASE OF QUASI-ROBUST CLUSTER STRUCTURE OF SET OF OBJECTS

D.A. VIATTCHENIN, A.V. DAMARATSKI

Abstract

A method of clustering of objects for varying in an interval attributes values in cases of the robust or quasi-robust cluster structure of the set of objects is proposed.

Литература

1. *Вятченин Д.А.* // Докл. БГУИР. 2009. № 6. С. 91–98.
2. *Bezdek J.C.* Pattern Recognition with Fuzzy Objective Function Algorithms. New York, 1981.
3. *Вятченин Д.А.* // Искусственный интеллект. 2007. № 3. С. 205–216.
4. *Viattchenin D.A.* // Control & Cybernetics. 2004. Vol. 33. P.323–340.
5. *Turksen B.* // Fuzzy Sets and Systems. 1986. Vol. 20. P. 191–210.
6. *Viattchenin D.A.* // Journal of Uncertain Systems. 2009. Vol. 3. P. 64–80.
7. *Ju H., Yan X.* // Fuzzy Information and Engineering. Berlin: Springer-Verlag, 2007. P. 875–883.
8. *Аверкин А.Н., Батыршин И.З., Блишун А.Ф. et al.* Нечеткие множества в моделях управления и искусственного интеллекта. М., 1986.
9. *Sato-Ilic M., Jain L.C.* Innovations in Fuzzy Clustering. Heidelberg, 2006.
10. *Вятченин Д.А.* // Вести Института современных знаний. 2008. № 4. С. 95–101.
11. *Вятченин Д.А., Доморацкий А.В., Новиков Д.И., Юодялис А.В.* // Материалы конференции ИММОД-2009. СПб., 2009. С. 109–113.