УДК 510.22+519.237.8

АНАЛИЗ УСТОЙЧИВОСТИ КЛАСТЕРНОЙ СТРУКТУРЫ В ЗАДАЧАХ НЕСТАЦИОНАРНОЙ КЛАСТЕРИЗАЦИИ

Д.А. ВЯТЧЕНИН

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

Поступила в редакцию 24 февраля 2009

Формулируется задача определения устойчивости кластерной структуры при автоматической классификации объектов с варьирующимися в интервале значениями признаков и представлен метод ее решения, основанный на эвристическом алгоритме возможностной кластеризации.

Ключевые слова: возможностная кластеризация, распределение по нечетким кластерам, типичная точка, динамические признаки, устойчивость кластерной структуры.

Введение

В задачах кластеризации, именуемых также задачами автоматической классификации, необходимо принимать во внимание, что признаки классифицируемых объектов могут изменять свои значения с течением времени или при наличии внешних воздействий. Объекты, характеризующиеся подобными признаками, называются динамическими [1], нечеткими [2], либо объектами с динамическими признаками [3]. Естественно, что кластерная структура исследуемой совокупности, состоящей из таких объектов, также является динамической и зависит от значений признаков и их количества в момент классификации.

Задачи автоматической классификации динамических объектов условно объединяются в класс задач нестационарной кластеризации, который, в свою очередь, включает в себя ряд подклассов задач, таких, например, как задачи структурно-классификационного прогнозирования многомерных динамических объектов [1], задачи динамической кластеризации [4] и задачи кластеризации так называемых многонаправленных данных [5, 6]. Следует отметить, что основой существующих методов решения всех перечисленных типов задач являются методы нечеткой кластеризации [7], 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 $x_i \in X \quad \forall i = 1, \ldots, n$, принадлежит нечеткому кластеру $A^l, l = 1, \ldots, c$. Разновидностью методов нечеткой кластеризации являются методы возможностной кластеризации [8], специфика которых заключается в том, что структура, образуемая нечеткими кластерами, удовлетворяет условию возможностного разбиения, являющегося менее жестким, чем условия нечеткого разбиения или нечеткого покрытия, и значения принадлежности интерпретируются как степени типичности объекта x_i нечеткому кластеру, а функция принадлежности интерпретируется как функция распределения возможностей.

В статье на содержательном уровне формулируется задача построения устойчивой кластерной структуры совокупности объектов с варьирующимися в интервале значениями

признаков и представлен метод определения устойчивости кластерной структуры, в основе которого лежит алгоритм построения распределения объектов по нечетким α-кластерам.

Постановка задачи

В задачах кластеризации данные об исследуемой совокупности традиционно представлены матрицей $X_{n\times m_1}=[\hat{x}_i^{t_1}]$, $i=1,\ldots,n$, $t_1=1,\ldots,m_1$, где x_i, $i=1,\ldots,n$, — объекты исследуемой совокупности X, а $\hat{x}^{t_1}$, $t_1=1,\ldots,m_1$, — значения признаков объектов $x_i\in X$, каждый из которых, таким образом, представляет собой точку в m_1-мерном признаковом пространстве. Данные об объектах исследуемой совокупности именуются многонаправленными, если они содержат также информацию о значениях признаков объектов в различных состояниях — в случае, когда данные представляют собой описание вида "объект–признак–состояние", так что каждый признак $\hat{x}^{t_1}$, $t_1=1,\ldots,m_1$ описывается m_2-мерным вектором значений $\hat{x}^{t_1}=(\hat{x}^{t_1(1)},\ldots,\hat{x}^{t_1(m_2)})$ и каждый объект x_i, $i=1,\ldots,n$, исследуемой совокупности X может быть представлен матрицей вида $X_{(i)m_1\times m_2}=[\hat{x}_i^{t_1(t_2)}]$, $t_1=1,\ldots,m_1$, $t_2=1,\ldots,m_2$ [3], данные именуются тринаправленными [5]. В [5] задача кластеризации многонаправленных данных состоит в построении нечеткого c-разбиения исследуемой совокупности n объектов $X=\{x_1,\ldots,x_n\}$ на априори известное число классов, и ее решение основано на многокритериальной оптимизации, а в [6] задача состоит в построении возможностного разбиения объектов $x_i\in X$, $i=1,\ldots,n$, на заданное c число классов, однако основой предложенного метода является построение матрицы слабой нечеткой толерантности, представляющей геометрическую структуру исследуемой совокупности объектов.

В задачах динамической кластеризации предполагается, что каждый объект x_i, $i=1,\ldots,n$, исследуемой совокупности X характеризуется m_1-мерным вектором признаков, измеренных в дискретные моменты времени $t_2\in\{1,\ldots,m_2\}$, так что объект $x_i\in X$ в момент времени t_2 представляется точкой $\hat{x}_i^{t_2}=(\hat{x}_i^{1(t_2)},\ldots,\hat{x}_i^{t_1(t_2)},\ldots,\hat{x}_i^{m_1(t_2)})$, причем число объектов не предполагается фиксированным, так как при анализе кластерной структуры рассмотрению подлежат только объекты, которые фиксировались в момент, предшествовавший текущему, и задача заключается в обнаружении временного изменения кластерной структуры с ее последующей коррекцией [4].

В свою очередь, в задачах структурно-классификационного прогнозирования многомерных динамических объектов [1] число классов c, как и число объектов исследуемой совокупности n, предполагается известным, и задача состоит в прогнозировании класса каждого объекта $x_i\in X$, описанного в виде $\hat{x}_i^{t_2}=(\hat{x}_i^{1(t_2)},\ldots,\hat{x}_i^{t_1(t_2)},\ldots,\hat{x}_i^{m_1(t_2)})\ \ \forall t_2=1,\ldots,m_2$, в момент времени m_2+1.

Вместе с тем признаки $\hat{x}^{t_1}$, $t_1=1,\ldots,m_1$ объектов $x_i\in X$ могут принимать не какое-то конкретное значение в некоторый дискретный момент времени $t_2\in\{1,\ldots,m_2\}$, и, таким образом, описываться вектором значений $\hat{x}^{t_1}=(\hat{x}^{t_1(1)},\ldots,\hat{x}^{t_1(m_2)})$, а 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 $[\hat{x}_i^{t_1\min},\hat{x}_i^{t_1\max}]$. В подобной ситуации кластерная структура исследуемой совокупности X также является динамической, однако в ряде задач оказывается важным установить, образуют ли классифицируемые объекты устойчивую кластерную структуру. В [4] выделяются следующие основные типы динамических изменений кластерной структуры, образуемой объектами исследуемой совокупности с течением времени:

1) образование новых кластеров;

2) слияние кластеров;

3) расщепление или дробление кластеров;
4) элиминация или исчезновение кластеров;
5) дрейф прототипов кластеров.

Как отмечается в [4], первые четыре типа изменений кластерной структуры носят скачкообразный характер, тогда как пятый тип изменений кластерной структуры носит непрерывный, и, как правило, латентный характер. Вместе с тем изменение координат прототипов кластеров указывает, с одной стороны, на потенциальное изменение числа кластеров, а с другой — на геометрическую форму, и, как следствие, свойства кластеров рассматриваемой структуры, что имеет большое значение в практических приложениях. Следует отметить, что подобные изменения кластерной структуры могут иметь место не с течением времени, а зависеть от значений признаков объектов исследуемой совокупности, которые могут изменяться не только с течением времени, но и в зависимости от характера внешней среды.

Содержательно задача построения устойчивой кластерной структуры, в общем, может быть сформулирована 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 $[\hat{x}_i^{t_1 \min}, \hat{x}_i^{t_1 \max}]$.

Для построения устойчивой кластерной структуры исследуемой совокупности объектов в первую очередь необходимо установить тип динамических изменений кластерной структуры. Если при изменении в соответствующем интервале $[\hat{x}_i^{t_1 \min}, \hat{x}_i^{t_1 \max}]$ 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 $\{\bar{\tau}^1,\ldots,\bar{\tau}^c\}$, то структура, образуемая кластерами $\{A^1,...,A^c\}$, будет именоваться устойчивой, если с изменением значений признаков объектов число c кластеров $\{A^1,...,A^c\}$ не изменяется, но изменяются координаты их прототипов $\{\bar{\tau}^1,\ldots,\bar{\tau}^c\}$, то соответствующая кластерная структура будет именоваться квазиустойчивой, а если при изменении значений признаков наблюдаемых объектов $x_i \in X$ изменяется число c кластеров, то кластерная структура будет именоваться неустойчивой. Первые четыре из перечисленных выше типов динамических изменений обуславливаются неустойчивой кластерной структурой, а пятый тип изменений определяется квазиустойчивой кластерной структурой. Таким образом, задача заключается в определении типа кластерной структуры совокупности объектов с варьирующимися в интервале значениями признаков.

Эвристический метод возможностной кластеризации

В [9] предложен эвристический метод нечеткой кластеризации, заключающийся в построении распределения по априори задаваемому числу c нечетких α-кластеров, удовлетворяющих введенному определению, в силу чего базовая версия алгоритма, являющаяся процедурой прямой классификации, от аббревиатуры английских терминов direct — прямой и allotment among fuzzy clusters — распределение по нечетким кластерам, получила обозначение D-AFC(c)-алгоритма [10]. 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 об 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

$$\mu_{li} = \begin{cases} \mu_{A^l}(x_i), & x_i \in A^l_\alpha \\ 0, & \text{иначе} \end{cases}, \qquad (1)$$

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 (1), показывает степень сходства i-го объекта множества X с типичной точкой τ^l соответствующего нечеткого α-кластера. Следует отметить, что в соответствии с принятой в [9] терминологией, функция принадлежности μ_{li} в смысле выражения (1) интерпретируется как степень типичности i-го объекта для l-го нечеткого α-кластера, и термин "типичность" использовался в [9] для определения содержательного смысла функции принадлежности (1) независимо от работы Р. Кришнапурама и Дж.М. Келлера [8]. В свою очередь, если условие

$$\mu_{li} \geq 0,\ \sum_{l=1}^{c} \mu_{li} > 0, l = 1,\ldots,c,\ i = 1,\ldots,n \qquad (2)$$

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 нечеткой толерантностью T, заданной на множестве объектов X, то это семейство является распределением множества классифицируемых объектов X по c нечетким α-кластерам. Условие (2), представляющее собой условие возможностного разбиения [8], требует, чтобы все объекты совокупности X были распределены по c нечетким α-кластерам $\{A^1_{(\alpha)},\ldots,A^c_{(\alpha)}\}$ с положительной степенью принадлежности μ_{li}, $l = 1,\ldots,c$, $i = 1,\ldots,n$. Таким образом, распределение по нечетким α-кластерам представляет собой частный случай возможностного разбиения, и соответствующая процедура представляет собой эвристический алгоритм возможностной кластеризации [10].

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 распределений $R^\alpha_z(X) \in B(c)$ критерия, определяющего качество каждого распределения:

$$F(R^\alpha_z(X), \alpha) = \sum_{l=1}^{c} \frac{1}{n_l} \sum_{i=1}^{n_l} \mu_{li} - \alpha c, \qquad (3)$$

где $n_l = \text{card}(A^l_\alpha)$ — мощность носителя нечеткого множества $A^l_{(\alpha)} \in R^\alpha_z(X)$, $l \in \{1,\ldots,c\}$, $\alpha \in (0,1]$, так что критерий (3) определяет среднюю суммарную принадлежность объектов исследуемой совокупности X нечетким α-кластерам $\{A^1_{(\alpha)},\ldots,A^c_{(\alpha)}\}$ распределения $R^\alpha_z(X)$ за вычетом величины αc, регуляризующей число классов в $R^\alpha_z(X)$, и наилучшему из множества $B(c)$ распределению объектов $R^*(X)$ соответствует максимальное значение (3), так что решение задачи состоит в построении распределения, удовлетворяющего условию

$$R^*(X) = \arg \max_{R^\alpha_z(X) \in B(c)} F(R^\alpha_z(X), \alpha). \qquad (4)$$

Результатом работы D-AFC(c)-алгоритма будет не только распределение $R^*(X)$ объектов исследуемой 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 [11] предложена ее модификация, использующая транзитивное замыкание нечеткой толерантности, в силу чего от аббревиатуры выражения transitive closure — предложенная в [11] процедура получила условное обозначение D-AFC-TC-алгоритма. Так как транзитивное замыкание нечеткой толерантности представляет собой нечеткую эквивалентность, разбивающую предметную область на непересекающиеся классы, то 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 $0 < \alpha_0 < \ldots < \alpha_\ell < \ldots < \alpha_Z = 1$ на основе вычисления скачка значений порога α определяется такое α_ℓ, которому соответствует некоторое неизвестное "естественное" число нечетких α-кластеров c. Помимо того, что D-AFC-TC-алгоритм отыскивает априори неизвестное число c нечетких α-кластеров, отличающими его от D-AFC(c)-алгоритма особенностями являются, во-первых, то, что для D-AFC-TC-алгоритма матрицей исходных данных является матрица "объект–признак" и для решения задачи классификации используются как критерий (5), так и некоторая метрика $d(x_i, x_j)$, в качестве которой могут выступать относительное обобщенное расстояние Хемминга, относительное евклидово расстояние или относительная евклидова норма [12], а во-вторых, то, что результатом работы D-AFC-TC-алгоритма будут также координаты прототипов $\{\bar{\tau}^1, \ldots, \bar{\tau}^c\}$ нечетких α-кластеров $\{A^1_{(\alpha)}, \ldots, A^c_{(\alpha)}\}$ распределения $R^*(X)$.

Метод определения устойчивости кластерной структуры

Для определения типа кластерной структуры предполагается, как это рассматривается в [13], классифицировать исследуемую совокупность объектов для предельных значений их признаков, т.е. исходными данными для решения задачи классификации является матрица значений признаков $X^{\min}_{n\times m_1} = [x_i^{t_1 \min}]$, $i = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$, где $x_i^{t_1 \min}$ соответствует наименьшему значению t_1-го признака для i-го объекта, и матрица значений признаков $X^{\max}_{n\times m_1} = [x_i^{t_1 \max}]$, $i = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$, где $x_i^{t_1 \max}$ соответствует наибольшему значению t_1-го признака для i-го объекта. Особенности D-AFC-TC-алгоритма, заключающиеся в автоматическом определении числа c нечетких α-кластеров вместе с их прототипами, позволяют использовать его для решения стоящей задачи. Таким образом, предлагаемый метод определения типа кластерной структуры заключается в построении матриц $X^{\min}_{n\times m_1} = [x_i^{t_1 \min}]$ и $X^{\max}_{n\times m_1} = [x_i^{t_1 \max}]$ с их последующей обработкой D-AFC-TC-алгоритмом при выбранном исследователем способе нормировки и расстоянии между нечеткими множествами с последующим сопоставлением числа c нечетких α-кластеров в распределениях $R^*(X)$, координат их прототипов и значений принадлежности объектов, полученных при обработке каждой из двух матриц. В отличие от предлагаемого подхода, изложенный в [13] метод состоит в построении нечеткого c-разбиения каждой из матриц с последующим рассмотрением изменений значений принадлежностей объектов классам нечеткого c-разбиения для задаваемого исследователем числа c классов, что является сильным допущением.

Иллюстративный пример

Эффективность предложенного подхода к определению устойчивости кластерной структуры целесообразно проиллюстрировать на примере, и для проведения вычислительного

эксперимента были выбраны представленные в табл. 1 тестовые данные о 8 объектах, описываемых тремя признаками, каждый из которых принимает значение в интервале [13].

Таблица 1. **Тестовые данные М. Сато-Илик и Л. Джейна**

Номер объекта	Номера признаков		
	1	2	3
1	[10, 10]	[1, 1]	[2, 2]
2	[9, 9]	[3, 3]	[4, 4]
3	[13, 13]	[3, 3]	[2, 2]
4	[14, 14]	[4, 4]	[5, 5]
5	[4, 8]	[11, 11]	[2, 12]
6	[6, 10]	[9, 9]	[1, 8]
7	[2, 11]	[10, 10]	[1, 11]
8	[3, 9]	[8, 8]	[2, 9]

Обозначая объекты символами x_i, $i=1,\ldots,8$, а признаки — соответственно символами $\hat{x}^{t_1}$, $t_1=1,\ldots,3$, были получены две матрицы "объект–признак" $X_{8\times3}^{\min}=[\hat{x}_i^{t_1\min}]$ и $X_{8\times3}^{\max}=[\hat{x}_i^{t_1\max}]$, обработанные с помощью нормализации

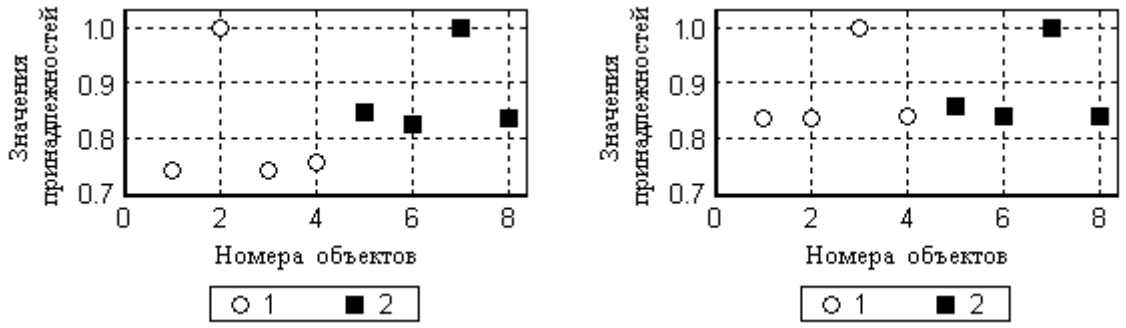
$$x_i^{t_1}=\frac{\hat{x}_i^{t_1}}{\max\limits_i \hat{x}_i^{t_1}}\ ,\ i=1,\ldots,n,\ \ t_1=1,\ldots,m_1, \qquad (5)$$

вследствие чего каждый объект может интерпретироваться как нечеткое множество на универсуме признаков с функцией принадлежности $\mu_{x_i}(x^{t_1})$, $i=1,\ldots,n$, с последующим использованием в D-AFC-TC-алгоритме относительного евклидова расстояния [12] между нечеткими множествами:

$$e(x_i,x_j)=\sqrt{\frac{1}{m_1}\sum_{t=1}^{m_1}\left(\mu_{x_i}(x^{t_1})-\mu_{x_j}(x^{t_1})\right)^2}\ ,\ i=1,\ldots,n,\ \ t_1=1,\ldots,m_1, \qquad (6)$$

и применением операции дополнения для построения нечеткой толерантности [11].

Значения принадлежностей объектов, полученные в результате обработки матриц $X_{8\times3}^{\min}=[\hat{x}_i^{t_1\min}]$ и $X_{8\times3}^{\max}=[\hat{x}_i^{t_1\max}]$ D-AFC-TC-алгоритмом, приведены на рисунке.



а *б*

Значения принадлежностей объектов двум классам при кластеризации по наименьшим (*а*) и наибольшим (*б*) значениям динамических признаков

В свою очередь, в табл. 2 приведены координаты прототипов двух нечетких α-кластеров, полученных в обоих случаях распределений $R^*(X)$.

Поскольку число классов в обоих распределениях оказывается одинаковым, а координаты прототипа нечеткого α-кластера, соответствующего второму классу, в зависимости от вида обрабатываемой матрицы исходных данных оказываются различными, очевидно, что кластерная структура является квазиустойчивой.

Таблица 2. **Координаты прототипов нечетких α-кластеров**

Номер класса	При обработке матрицы наименьших значений признаков			При обработке матрицы наибольших значений признаков		
	номера признаков			номера признаков		
	1	2	3	1	2	3
1	11,5	2,75	3,25	11,5	2,75	3,25
2	3,75	9,5	1,5	9,5	9,5	10

Следует отметить, что для первых четырех объектов исследуемой совокупности, как следует из табл. 1, значения признаков не изменяются, как и не изменяются координаты прототипа нечеткого α-кластера, соответствующего первому классу, в котором оказались указанные объекты, при обработке как матрицы $X_{8\times3}^{\min} = [\mathbf{x}_i^{1\min}]$, так и матрицы $X_{8\times3}^{\max} = [\mathbf{x}_i^{1\max}]$, однако для первого класса типичной точкой, т.е. объектом, обладающим наибольшей степенью принадлежности в смысле (1) к тому или иному классу, при обработке матрицы $X_{8\times3}^{\min} = [\mathbf{x}_i^{1\min}]$ является второй объект, а при обработке матрицы $X_{8\times3}^{\max} = [\mathbf{x}_i^{1\max}]$ — третий. Данное обстоятельство, как и существенное изменение значений принадлежности объектов первого класса, обусловлено, в первую очередь, выбором (5) для нормировки исходных данных. Например, используя для нормировки исходных данных унитаризацию [10], а в качестве расстояния вместо (6) — относительную евклидову норму [12], в обоих случаях также было получено распределение по двум нечетким α-кластерам, но при обработке матрицы $X_{8\times3}^{\min} = [\mathbf{x}_i^{1\min}]$ типичными точками нечетких α-кластеров оказались объекты x_2 и x_7, а при обработке матрицы $X_{8\times3}^{\max} = [\mathbf{x}_i^{1\max}]$ — объекты x_3 и x_8 соответственно, причем, как и в рассмотренном случае, имел место дрейф прототипа второго класса.

Для детального анализа результатов эксперимента с целью сравнения представляется целесообразным привести результаты, представленные в [13], где исходные данные были разбиты на два класса FANNY-алгоритмом нечеткой кластеризации, предложенным Л. Кофманом и П.Дж. Рауссеу в [14].

Таблица 3. **Результаты обработки тестовых данных FANNY-алгоритмом**

Номер объекта	Значения принадлежностей объектов классам			
	при обработке матрицы наименьших значений признаков		при обработке матрицы наибольших значений признаков	
	1	2	1	2
1	0,86	0,14	0,87	0,13
2	0,82	0,18	0,82	0,18
3	0,90	0,11	0,88	0,12
4	0,84	0,16	0,76	0,24
5	0,09	0,91	0,15	0,85
6	0,18	0,82	0,15	0,85
7	0,09	0,91	0,12	0,88
8	0,11	0,89	0,13	0,87

Сравнительный анализ результатов, представленных на рисунке и в табл. 3 демонстрирует их некоторое сходство, однако из представленных в табл. 3 матриц нечетких c-разбиений однозначного вывода о типе кластерной структуры сделать нельзя, что наглядно демонстрирует преимущество предлагаемого подхода.

Заключение

Анализ приведенных результатов наглядно демонстрирует, помимо указанного по сравнению с методом М. Сато-Илик и Л. Джейна [13] преимущества, существенную особенность предлагаемого подхода, заключающуюся в возможности рассмотрения вместо изменения координат прототипов кластеров изменения номеров типичных точек нечетких α-кластеров распределений $R^*(X)$, полученных при обработке матриц $X_{n\times m_1}^{\min} = [\mathbf{x}_i^{1\min}]$ и

$X^{\max}_{n\times m_1} = [\hat{x}_i^{t_1 \max}]$. Если в полученных распределениях число классов c остается неизменным и не изменяются номера типичных точек нечетких α-кластеров — элементов полученных при обработке матриц $X^{\min}_{n\times m_1} = [\hat{x}_i^{t_1 \min}]$ и $X^{\max}_{n\times m_1} = [\hat{x}_i^{t_1 \max}]$ распределений $R^*(X)$, то даже в случае дрейфа прототипов кластеров кластерная структура может считаться устойчивой, а дрейф прототипов — несущественным.

Направление дальнейших исследований определяется необходимостью обобщения предложенного в [6] подхода, позволяющего определять меру различия между динамическими объектами, на случай интервально-значных нечетких множеств с целью построения матрицы слабой нечеткой толерантности, представляющей геометрическую структуру исследуемой совокупности. Если в результате проведенного с помощью предложенного подхода анализа кластерная структура исследуемой совокупности окажется устойчивой или квазиустойчивой, то матрица слабой нечеткой толерантности должна разбиваться D-AFC(c)-алгоритмом на определенное в результате анализа число c классов. В случае если кластерная структура оказывается неустойчивой, то определяются значения наименее возможного c_* и наиболее возможного c^* числа областей признакового пространства $\Re^m$, в которых отображаются кластеры при различных значениях признаков, принимаемых объектами совокупности X, после чего матрица слабой нечеткой толерантности обрабатывается D-AFC(c)-алгоритмом для всех $c \in [c_*, c^*]$ с определением оптимального числа c на основе вычисления некоторого показателя валидности числа нечетких α-кластеров, разработка которого также представляет собой одно из направлений дальнейших исследований.

ANALYSIS OF THE CLUSTER STRUCTURE ROBUSTNESS IN NONSTATIONARY CLUSTERING PROBLEMS

D.A. VIATTCHENIN

Abstract

The problem of explaining of robustness of cluster structure in automatic classification of objects for varying in an interval attributes values is formulated and a method for the problem solving based on a heuristic algorithm of possible clustering is presented.

Литература

1. *Дорофеюк Ю.А., Дорофеюк А.А.* // Искусственный интеллект. 2006. № 2. С. 138–141.
2. *Žák L.* // Archivum Mathematicum. 2002. Vol. 38. P. 37–50.
3. *Вятченин Д.А.* // Вестник Военной академии Республики Беларусь. 2005. № 3. С. 32–37.
4. *Гимаров В.А., Дли М.И., Битюцкий С.Я.* // Нефтегазовое дело. 2004. № 2. С. 203–209.
5. *Sato M., Sato Y.* // Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems. 1994. Vol. 2. P. 127–142.
6. *Viattchenin D.A.* // J. of Uncertain Systems. 2009. Vol. 3. P. 64–80.
7. *Вятченин Д.А.* Нечеткие методы автоматической классификации. Минск, 2004.
8. *Krishnapuram R., Keller J.M.* // IEEE Trans. on Fuzzy Systems. 1993. Vol. 1. P. 98–110.
9. *Viattchenin D.A.* // Control & Cybernetics. 2004. Vol. 33. P. 323–340.
10. *Вятченин Д.А.* // Вести Института современных знаний. 2008. № 3. С. 205–216.
11. *Вятченин Д.А.* // Искусственный интеллект. 2007. № 3. С. 205–216.
12. *Кофман А.* Введение в теорию нечетких множеств. М., 1982.
13. *Sato-Ilic M., Jain L.C.* Innovations in Fuzzy Clustering: Theory and Applications. Heidelberg, 2006.
14. *Kaufman L., Rousseeuw P.J.* Finding Groups in Data: An Introduction to Cluster Analysis. N.Y., 1990.