УДК 519.237.8+510.22

МЕТОД МЯГКОЙ ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

© Вятченин Д.А.

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

e-mail: *viattchenin@mail.ru*

Abstract. The paper provides a new method of the interpretation of fuzzy clustering results which is allow put an automatic choosing of the threshold value that is a basis for detecting of sets of most informative elements of fuzzy classes. An effectiveness of the new method is illustrated on an example of fuzzy data fuzzy clustering results processing.

Введение

В последние годы значительно выросло количество публикаций, посвященных различным аспектам нечеткого подхода к решению задач автоматической классификации, или, иными словами, нечеткой кластеризации. Это обусловлено тем, что методы нечеткой кластеризации отличаются от традиционных «жестких» методов кластеризации, с одной стороны, высокой точностью, а с другой – содержательной осмысленностью результатов классификации. Как и в традиционных методах кластерного анализа, в рамках нечеткого подхода к решению задачи автоматической классификации выделяются эвристическое, оптимизационное и иерархическое направления, подробно рассматриваемые в работе [1]. Наиболее распространенным подходом к решению нечеткой модификации задачи автоматической классификации является оптимизационный подход [2], методы которого предусматривают нахождение оптимального, в смысле используемого критерия качества $Q(P(X))$, разбиения $P^*(X) = \{A^1, \ldots, A^c\}$ на заданное число c нечетких кластеров, описываемых функциями принадлежности μ_{li}, $l = 1, \ldots, c$, $i = 1, \ldots, n$, определенных на исследуемой совокупности объектов $X = \{x_1, ..., x_n\}$, так что задача нечеткой кластеризации заключается в нахождении экстремума целевой функции $Q(P(X))$, что в общем виде описывается формулой

$$Q(P(X)) \to \underset{P(X)\in\Pi}{extr}, \tag{1}$$

где Π – множество всех возможных нечетких разбиений $P(X)$ множества классифицируемых объектов X, при ограничениях, определяемых условием

$$\mu_{li} \geq 0, \quad \sum_{i=1}^{n} \mu_{li} = 1, \quad i = 1, \ldots, n, \quad l = 1, \ldots, c, \tag{2}$$

именуемым также условием нечеткого c-разбиения или нечеткого разбиения в смысле Распини [3], которое описывается матрицей $P_{c\times n} = [\,\mu_{li}]$, где $\mu_{li} = \mu_{A^l}(x_i)$ – значение принадлежности элемента $x_i \in X$ некоторому нечеткому кластеру $A^l \in \{A^1, \ldots A^c\}$.

Постановка проблемы, рассматриваемой в статье, заключается в отнесении объекта классифицируемой совокупности одному или нескольким классам в случае решения задачи классификации с помощью оптимизационных методов нечеткой кластеризации.

Анализ существующих результатов решения указанной проблемы, проведенный в рамках предпринятого исследования, демонстрирует, что в процессе интерпретации результатов нечеткой классификации в ряде случаев могут возникнуть трудности, связанные с проблемой однозначного отнесения объекта к тому или иному классу; кроме того, существующие подходы являются жесткими в том смысле, что при отнесении объекта к тому или иному классу, ассоциированному с соответствующим нечетким кластером полученного нечеткого c-разбиения, значения принадлежностей объектов нечетким кластерам элиминируются, так что применение нечеткой кластеризации с методологической точки зрения теряет смысл.

Целью исследования является обоснование метода «мягкой» интерпретации результатов нечеткой кластеризации, позволяющего, с одной стороны, отнести каждый объект исследуемой совокупности к наименьшему числу $\tilde{c}$, $1 \le \tilde{c} \le c$ нечетких кластеров нечеткого c-разбиения $P^*(X) = \{A^1, \ldots, A^c\}$, являющегося результатом классификации, а с другой – сохранить значения принадлежности μ_{li}, которые можно интерпретировать как степени обладания объектом $x_i \in X$ свойств класса, ассоциированного с нечетким кластером A^l, $l \in \{1, \ldots, c\}$ – элементом нечеткого c-разбиения $P^*(X)$, оптимального в смысле выбранного критерия качества $Q(P(X))$.

1. Основные методы интерпретации результатов нечеткой кластеризации

Наиболее простым и распространенным методом интерпретации результатов нечеткой классификации является дефаззификация матрицы $P_{c\times n} = [\,\mu_{li}]$ нечеткого c-разбиения по правилу максимального значения принадлежности [4]:

$$P_i^{MM} = \mathrm{e}_l \Leftrightarrow \mu_{li} > \mu_{ai},\ a = 1, 2, \ldots, c,\ a \neq l, \tag{3}$$

так что значениями принадлежности μ_{li}^{MM} матрицы $P_{c\times n}^{MM}$ являются числа 0 и 1, и принадлежность i-го объекта l-му классу определяется по формуле

$$\mu_{li}^{MM} = \begin{cases} 1, & \mu_{li} > \mu_{ai},\ a = 1, 2, \ldots, c, a \neq l \\ 0, & \text{в противном случае} \end{cases}. \tag{4}$$

Однако подобный подход является неприемлемым, если для некоторого объекта $x_i \in X$ его значения принадлежностей составляют $\mu_{li} = 1/c$, $l = 1, \ldots, c$.

С другой стороны, В. Педричем [5] для выявления структуры нечетких кластеров предлагается ввести два порога, один из которых определяется исследователем, а значение второго вычисляется, и которые позволяют выделять ядра нечетких кластеров. Так как значение μ_{li} выражает степень принадлежности i-го элемента l-му кластеру, то объект $x_i \in X$ может рассматриваться как элемент ядра нечеткого кластера A^l, если для некоторого порога φ имеет место $\mu_{li} > \varphi$. При рассмотрении для каждого $x_i \in X$ структурного отношения ко всем $A^l, l = 1, \ldots, c$, принимаются во внимание функции принадлежности точки x_i всем остальным нечетким кластерам,

за исключением рассматриваемого нечеткого кластера A^l. Поскольку число кластеров равно c, можно отметить, что любая точка $x_i \in X$, для которой $\mu_{li} = \frac{1}{c}, l = 1, \ldots, c$ не способствует выявлению сущности кластерной структуры. Мерой структурного свойства некоторой точки $x_i \in X$ служит показатель

$$\xi(x_i) = 1 - c^c \prod_{l=1}^{c} \mu_{li}, \tag{5}$$

то есть $\xi : X \to [0, 1]$, и из выражений (2) и (5) следует, что если для некоторого x_i и одного из кластеров A^l, $l = 1, \ldots, c$, будет $\mu_{li} = 1$, то значение показателя (5) в x_i максимально, то есть $\xi(x_i) = 1$, и точка x_i будет элементом ядра нечеткого кластера, если $\xi(x_i)$ превышает порог ϕ. Таким образом, понятие (ϕ, φ)-ядра нечеткого кластера A^l, $l = 1, \ldots, c$, может быть определено как подмножество $X^l_{\phi\varphi}$ универсума $X = \{x_1, ..., x_n\}$ при $\phi, \varphi \in [0, 1]$, содержащее множество $\left\{ x_i \in X \mid 1 - c^c \prod_{l=1}^{c} \mu_{li} \geq \phi,\ \mu_{li} \geq \varphi \right\}$, то есть

$$X^l_{\phi\varphi} = \{ x_i \in X \mid \xi(x_i) \geq \phi,\ \mu_{li} \geq \varphi \}, \tag{6}$$

так что множество ядер $X^l_{\phi\varphi}$, $l = 1, \ldots, c$, и резидуальное множество данных X^R, содержащее все оставшиеся элементы универсума X, связаны соотношением

$$X = \bigcup_{l=1}^{c} X^l_{\phi\varphi} \cup X^R, \tag{7}$$

где первая составляющая правой части является существенной для рассмотрения структурой, а вторая соответствует малосущественной структуре множества X.

Недостатком обоих подходов является утрата значений принадлежности объектов нечетким кластерам, позволяющая содержательно интерпретировать результаты кластеризации.

2. Понятие α-ядра нечеткого кластера

В свою очередь, концепция α-ядер нечетких кластеров, предложенная в работе [6] в рамках разработки методологии многостадийной нечеткой кластеризации, предполагает нахождение такого порога α, $\alpha \in (0, 1]$, чтобы выполнялось условие

$$X = \bigcup_{l=1}^{c} Supp(A^l(\alpha)), \tag{8}$$

где $X = \{x_1, \ldots, x_n\}$ – исследуемая совокупность объектов, α-ядра $A^l(\alpha)$, $l = 1, \ldots, c$, нечетких кластеров $A^l \in P$, $l = 1, \ldots, c$, для $\alpha \in (0, 1]$ представляют собой нечеткие множества уровня [7], определяемые как $A^l(\alpha) = \{(x_i, \mu^\alpha_{li}) \mid \mu^\alpha_{li} \geq \alpha\}$, $x_i \in X$, так что $A^l(\alpha) \subseteq A^l$, $\alpha \in (0, 1]$, $A^l \in \{A^1, \ldots, A^c\}$, а $Supp(A^l(\alpha))$ – носитель α-ядра $A^l(\alpha)$ нечеткого кластера $A^l \in P$, причем $Supp(A^l(\alpha)) = A^l_\alpha$, то есть носитель α-ядра нечеткого кластера $A^l \in P$, $l = 1, \ldots, c$ будет представлять собой α-срез

$A^l_\alpha = \{x_i \in X \mid \mu_{li} \geq \alpha\}$ [8] этого кластера при соответствующем значении α, а значения принадлежности объекта α-ядру нечеткого кластера определяется в соответствии с формулой

$$\mu^\alpha_{li} = \begin{cases} \mu_{li}, & x_i \in A^l_\alpha \\ 0, & x_i \notin A^l_\alpha \end{cases} . \tag{9}$$

Порог α должен выбираться так, чтобы каждый объект $x_i \in X$, $i = 1, \ldots, n$, принадлежал бы по меньшей мере одному α-ядру нечеткого кластера, и может вычисляться по формуле

$$\hat{\alpha} = \min_i \max_l \mu_{li}, \tag{10}$$

что, в свою очередь, позволяет сформулировать следующее утверждение.

Теорема. *Для нечеткого с-разбиения $P = \{A^1, \ldots, A^c\}$ исследуемой совокупности объектов $X = \{x_1, \ldots, x_n\}$ носители $\{A^1_\alpha, \ldots, A^c_\alpha\}$ α-ядер $\{A^1(\alpha), \ldots, A^c(\alpha)\}$ нечетких кластеров образуют покрытие исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ в том, и только в том случае, когда $\alpha \leq \hat{\alpha}$, $\alpha \in (0, 1]$, где $\hat{\alpha} \in (0, 1]$ вычисляется в соответствии с формулой (10).*

Доказательство. Если для некоторого $\alpha \in (0, 1]$ семейство множеств $C = \{A^1_\alpha, \ldots, A^c_\alpha\}$, являющихся носителями α-ядер нечетких кластеров $A^1, \ldots, A^c$, образует покрытие исследуемой совокупности $X = \{x_1, \ldots, x_n\}$, то каждый объект $x_i \in X$ является элементом по меньшей мере одного подмножества $A^l_\alpha \in C$. Если каждый объект $x_i \in X$ является элементом только одного подмножества $A^l_\alpha \in C$, то $\bigcap\limits_{l=1}^{c} A^l_\alpha = \varnothing$, а если каждый объект $x_i \in X$ является элементом более чем одного подмножества $A^l_\alpha \in C$, то будет иметь место $\bigcap\limits_{l=1}^{c} A^l_\alpha \neq \varnothing$. Так как $\bigcup\limits_{l=1}^{c} A^l_\alpha$ – наименьшее множество, содержащее все множества $\{A^1_\alpha, \ldots, A^c_\alpha\}$, и так как $\bigcap\limits_{l=1}^{c} A^l_\alpha \subseteq \bigcup\limits_{l=1}^{c} A^l_\alpha$, то очевидно, что $x_i \in \bigcup\limits_{l=1}^{c} A^l_\alpha$, $\forall i \in \{1, \ldots, n\}$.

В свою очередь, условие (10) можно переписать в следующем виде:

$$\hat{\alpha} = \bigwedge_i (A^1 \vee A^2 \vee \ldots \vee A^c), \tag{11}$$

где символом $\vee$ в теории нечетких множеств традиционно обозначается операция взятия максимума, а символом $\wedge$ – операция взятия минимума [8]. В силу ассоциативности операции $\vee$ имеет место $A^1 \vee A^2 \vee \ldots \vee A^c = A$, и в силу конечности множества X существует по крайней мере один элемент $x_i \in X$, для которого выполняется условие

$$\hat{\alpha} = \bigwedge_i \mu_A(x_i), \tag{12}$$

так что для некоторого $\bar{\alpha} > \hat{\alpha}$, $\alpha \in (0,1]$ будет иметь место $\mu_{A(\bar{\alpha})}(x_i) = 0$, и, как следствие, $x_i \notin Supp(A(\bar{\alpha}))$. Соответственно, будет иметь место $x_i \notin \bigcup_{l=1}^{c} A^l_{\bar{\alpha}}$, что доказывает корректность утверждения теоремы. □

Из теоремы вытекает ряд утверждений, которые, в силу их очевидности и ограниченности изложения, приводятся без доказательства.

Следствие 1. *Если $\alpha = \hat{\alpha}$, $\alpha \in (0,1]$, где значение $\hat{\alpha}$ вычисляется по формуле (10), то покрытие, образуемое носителями $\{A^1_\alpha, \ldots, A^c_\alpha\}$ α-ядер $\{A^1(\alpha), \ldots, A^c(\alpha)\}$ кластеров нечеткого c-разбиения $P(X) = \{A^1, \ldots, A^c\}$, минимально.*

Следствие 2. *Если в условии (8) имеет место равенство, то носители $\{A^1_\alpha, \ldots, A^c_\alpha\}$ α-ядер кластеров нечеткого c-разбиения образуют разбиение исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ на непересекающиеся множества.*

Следствие 3. *В случае нечеткого c-разбиения исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ на два класса при $\hat{\alpha} = 0.5$ в области пересечения носителей α-ядер обоих кластеров будет находиться по меньшей мере один объект $x_i \in X$, $i \in \{1, \ldots, n\}$.*

3. Иллюстративный пример

Эффективность предложенного метода можно проиллюстрировать на примере обработки нечеткого c-разбиения при $c = 3$, полученного с помощью предложенного М.-Ш. Янгом и Ч.-Х. Ко [9] FCN-алгоритма, совокупности 30 треугольных нечетких чисел $V_i = (m, a, b)_T$, $i = 1, \ldots, 30$, где m – модальное значение, а a и b – левый и правый коэффициенты нечеткости соответственно, представленных, как и их значения принадлежностей, в таблице 1.

Значения принадлежностей μ_{li}, $l = \overline{1,3}$, $i = 1, \ldots, 30$, объектов $V_1, \ldots, V_{30}$ классифицируемой совокупности нечетким кластерам A^l, $l = \overline{1,3}$, могут быть изображены в виде так называемой линейной диаграммы [1], представленной на рис. 1, где символом ∘ обозначены значения принадлежностей объектов нечеткому кластеру A^1, символом ■ – нечеткому кластеру A^2, а символом ▲ – нечеткому кластеру A^3.

Подобное представление результатов нечеткой кластеризации значительно усложняет их содержательную интерпретацию, особенно в случаях большого числа объектов исследуемой совокупности или числа нечетких кластеров в искомом нечетком c-разбиении, а также в случаях достаточно высоких значений принадлежности некоторых объектов нескольким нечетким кластерам одновременно.

В результате применения к матрице значений принадлежности $P_{c\times n} = [\mu_{li}]$, $l = \overline{1,3}$, $i = 1, \ldots, 30$, предложенного метода интерпретации, значение порога, позволяющего выделить α-ядра кластеров нечеткого c-разбиения, вычисляемого в соответствии с формулой (10), составило $\hat{\alpha} = 0.5255$, и значения принадлежностей объектов $\mu^{\hat{\alpha}}_{li}$, $l = \overline{1,3}$, $i = 1, \ldots, 30$, полученные по формуле (9), изображены на рис. 2.

Значения принадлежностей объектов α-ядрам нечетких кластеров демонстрируют их хорошую разделимость, что выражается в единственности положительного значения μ^{α}_{li} для каждого объекта V_i, $i = 1, \ldots, 30$ исследуемой совокупности.

Таблица 1. Совокупность 30 нечетких чисел и их значения принадлежностей нечетким кластерам

i	Параметры нечеткого числа			Значения принадлежностей		
	m	a	b	μ_{1i}	μ_{2i}	μ_{3i}
1	3.34	1.46	1.30	0.7836	0.1625	0.0538
2	9.56	0.27	1.00	0.9460	0.0437	0.0102
3	10.56	1.95	1.93	0.9680	0.0261	0.0057
4	10.89	0.56	1.17	0.9779	0.0182	0.0038
5	13.89	0.89	0.88	0.9866	0.0115	0.0018
6	14.78	0.12	1.21	0.9397	0.0529	0.0073
7	14.90	1.19	0.41	0.9511	0.0427	0.0061
8	15.67	1.82	0.90	0.8978	0.0904	0.0116
9	16.87	1.90	1.85	0.7412	0.2345	0.0241
10	17.45	1.79	1.95	0.6468	0.3236	0.0294
11	19.78	1.47	0.42	0.2929	0.6724	0.0345
12	20.67	1.34	1.10	0.1597	0.8124	0.0277
13	21.45	0.92	1.60	0.0760	0.9056	0.0183
14	22.34	0.04	1.58	0.0225	0.9698	0.0076
15	23.47	0.81	0.51	0.0041	0.9940	0.0018
16	24.67	0.14	1.09	0.0034	0.9942	0.0023
17	25.78	0.39	1.51	0.0189	0.9628	0.0181
18	26.45	1.61	0.92	0.0254	0.9473	0.0271
19	28.34	1.95	0.12	0.0556	0.8451	0.0992
20	32.29	1.66	1.64	0.0709	0.4035	0.5255
21	32.77	0.63	0.47	0.0658	0.3517	0.5824
22	34.88	1.08	0.66	0.0360	0.1527	0.8112
23	35.45	1.48	1.26	0.0274	0.1101	0.8624
24	35.88	1.79	0.16	0.0248	0.0978	0.8773
25	38.88	0.66	0.64	0.0004	0.0014	0.9980
26	40.25	0.52	1.71	0.0011	0.0034	0.9953
27	40.47	1.95	0.15	0.0006	0.0017	0.9976
28	43.56	0.92	0.63	0.0164	0.0412	0.9423
29	43.98	1.74	1.69	0.0195	0.0482	0.9321
30	45.77	1.71	0.79	0.0315	0.0735	0.8949

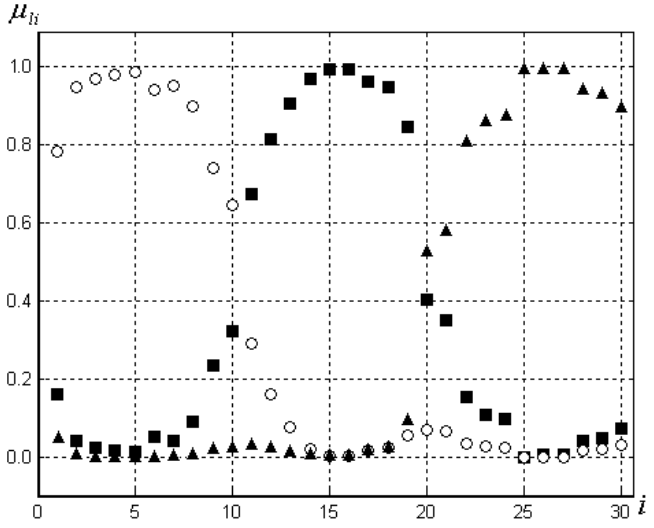

Рис. 1. Диаграмма значений принадлежности объектов нечетким кластерам

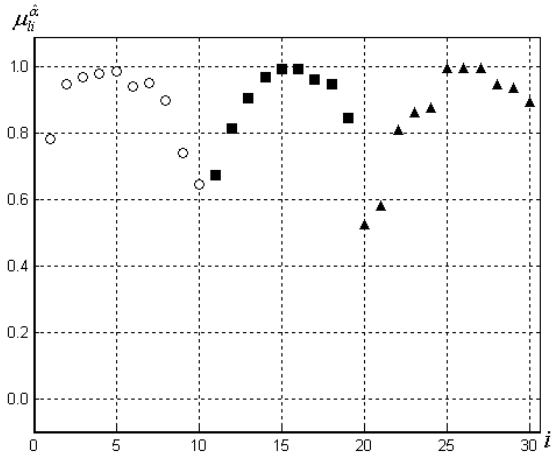

Рис. 2. Диаграмма значений принадлежности объектов α-ядрам нечетких кластеров

Заключение

Результатом проведенного исследования является метод выделения α-ядер нечетких кластеров, представляющих собой наиболее информативные, с точки зрения интерпретации результатов нечеткой классификации, нечеткие подмножества объектов исследуемой совокупности.

Предложенная концепция α-ядер нечетких кластеров обладает большей гибкостью, чем правило максимального значения принадлежности (3) и концепция (ϕ, φ)-ядер нечетких кластеров [5], так как сохраняет значения принадлежности μ_{li}^{α} объектов, что является немаловажным для содержательной интерпретации результатов кластеризации; кроме того, определение порога $\hat{\alpha}$ зависит только от значений принадлежности μ_{li}, $l = 1, \ldots, c$, $i = 1, \ldots, n$, в матрице $P_{c\times n} = [\mu_{li}]$ нечеткого c-разбиения, и не зависит от классифицируемых объектов и их признаков, формы и других характеристик нечетких кластеров.

Необходимо указать, что в случаях, когда объем исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ достаточно велик, носители $\{A_{\alpha}^1, \ldots, A_{\alpha}^c\}$ α-ядер $\{A^1(\alpha), \ldots, A^c(\alpha)\}$ кластеров нечеткого c-разбиения $P(X) = \{A^1, \ldots, A^c\}$ могут рассматриваться как множества объектов, подлежащие дальнейшей классификации. Данное обстоятельство является основополагающим при последовательном применении методов нечеткой кластеризации к большим массивам данных [6], и выделение α-ядер нечетких кластеров позволяет обрабатывать данные в полностью автоматическом режиме.

Список литературы

1. *Вятченин Д.А.* Нечеткие методы автоматической классификации. – Мн.: УП «Технопринт», 2004. – 219 с.
2. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition / Höppner F., Klawonn F., Kruse R., Runkler T. – Chichester: Wiley Intersciences, 1999. – 289 p.

3. *Заде Л.А.* Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер / Под ред. Дж. Вэн Райзина; пер с англ.; под ред. Ю.И. Журавлева. – М: Мир, 1980. – С. 208-247.
4. *Bensaid A.M., Hall L.O., Bezdek J.C., Clarke L.P.* Partially supervised clustering for image segmentation // Pattern Recognition. – 1996. – Vol. 29. – P. 859-871.
5. *Pedrycz W.* Fuzzy sets in pattern recognition: methodology and methods // Pattern Recognition. – 1990. – Vol.23. – P. 121-146.
6. *Viattchenin D.A.* Methodological Aspects of Fuzzy Clustering Application to Data Analysis Problems // Proceedings of the Sixth ISTC Scientific Advisory Committee Seminar (Moscow, Russia, September 15-17, 2003), Vol. I. – Moscow: Russian Academy of Sciences, 2003. – P. 286-293.
7. *Radecki T.* Level fuzzy sets // Journal of Cybernetics. – 1977. – Vol. 7. – P. 189-198.
8. *Кофман А.* Введение в теорию нечетких множеств / Пер. с фр. В.Б. Кузьмина; под ред. С.И. Травкина. – М.: Радио и связь, 1982. – 432 с.
9. *Yang M.-S., Ko C.-H.* On a class of fuzzy c-numbers clustering procedures for fuzzy data // Fuzzy Sets and Systems. – 1996. – Vol. 84. – P. 49-60.

Статья поступила в редакцию 25.04.2008