# ОБОБЩЕННАЯ МЕТОДОЛОГИЯ ОПРЕДЕЛЕНИЯ ЧИСЛА КЛАСТЕРОВ В НЕЧЕТКОМ *C*-РАЗБИЕНИИ ПРИ КЛАССИФИКАЦИИ МОДЕЛЕЙ

**К. М. Садовская (Минск, Беларусь)**

Как отмечается в работе [6], «в настоящее время особую значимость начинают приобретать вопросы оценивания и анализа качества моделей, классификации, упорядочения и обоснованного выбора состава и структур моделей и полимодельных комплексов, управления их качеством». Классификация моделей некоторого процесса или объекта является необходимым этапом исследования, определяющим выбор модели, наиболее адекватной условиям поставленной задачи. Если $X = \{x_1,..., x_n\}$ – множество $n$ моделей, описывающих некоторый процесс или объект $\chi$, и каждая модель $x_i \in X, i = 1,\ldots, n$ характеризуется $m$ признаками, то задача классификации моделей состоит в разбиении множества $X$, данные о котором представлены в виде матрицы

$$X_{n\times m} = \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^m \\ x_2^1 & x_2^2 & \ldots & x_2^m \\ \ldots & \ldots & \ldots & \ldots \\ x_n^1 & x_n^2 & \ldots & x_n^m \end{pmatrix}, \qquad (1)$$

именуемой матрицей «объект–свойство» [4], где $x_i^t$ представляет собой значение $t$-го признака у $i$-го объекта множества $X$, на заданное или нет число $c$ классов.

В последние годы значительно выросло количество публикаций, посвященных различным аспектам нечеткого подхода к решению задач классификации. Это обусловлено тем, что, как отмечал профессор Л. А. Заде, «большинство реальных кластеров размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен» [3]. Наибольшее распространение среди нечетких методов распознавания образов получили нечеткие методы автоматической классификации, подробно рассматриваемые в работе [1], которые объединяются в эвристическое, оптимизационное и иерархическое направления.

Оптимизационные методы нашли наибольшее распространение при решении задач нечеткой кластеризации. При их использовании целью решения задачи является разбиение исследуемой совокупности объектов на семейство $c$ нечетких множеств, называемых нечеткими кластерами. Нечеткие множества $A^l, l = 1,\ldots, c$, определенные на множестве классифицируемых объектов $X = \{x_1,..., x_n\}$, с соответствующими функциями принадлежности $\mu_{li}, l = 1,\ldots, c, i = 1,\ldots, n$, образуют нечеткое $c$-разбиение $P = \{A^1,\ldots, A^c\}$ на заданное число $c$ классов, если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^{c} \mu_{li} = 1$, так что задача нечеткой кластеризации заключается в нахождении экстремума некоторого функционала $Q(P)$ на множестве П всех нечетких $c$-разбиений $P$ исследуемой совокупности объектов $X = \{x_1,..., x_n\}$, что описывается формулой $Q(P) \to \underset{P\in\Pi}{extr}$. Решением задачи классификации будет нечеткое разбиение $P*$ на заданное число классов $c$, соответствующее, в зависимости от вида целевой функции, наименьшему либо, напротив, наибольшему значению $Q(P)$. Поскольку в качестве входного параметра во всех существующих оптимизационных нечетких методах автоматической классификации должно задаваться число нечетких кластеров $c$, которое может быть и неизвестным исследователю, возникает проблема определения наиболее

адекватного числа $c$ классов в искомом нечетком $c$-разбиении $P*$, для решения которой, в свою очередь, различными исследователями был предложен ряд показателей, вычисление которых для различного числа нечетких кластеров помогает определить наиболее «приемлемое» значение $c$, т. е. наиболее «естественное» число нечетких кластеров, на которое «расслаивается» исследуемая совокупность объектов. Для всех показателей числа классов в нечетком $c$-разбиении решение задачи определения оптимального числа классов в искомом нечетком разбиении определяется общим выражением

$$\underset{c}{extr}\left(V_c(P)\right), c = 2, \ldots, n-1, \qquad (2)$$

где символом $V_c(P)$ обозначен какой-либо показатель.

В работе [2] было предложено объединить алгоритм, вычисляющий нечеткое $c$-разбиение, оптимальное в смысле используемого функционала $Q(P)$, и процедуру вычисления соответствующего этому алгоритму показателя оптимальности числа нечетких классов в нечетком разбиении, в рамках одной процедуры. Процедура, объединяющая $FCM$-алгоритм, отыскивающий нечеткое $c$-разбиение, оптимальное в смысле функционала Дж. Данна и Дж. Беждека [7]

$$Q(P) = \sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i - \tau^l\right\|^2, \qquad (3)$$

и процедуру вычисления некоторого показателя оптимальности числа классов в нечетком $c$-разбиении, ряд которых рассмотрен в работе [2], для различных значений числа классов $c$, получила название $FCM-CV$-алгоритма. Параметрами $FCM-CV$-алгоритма являются показатель нечеткости классификации $\gamma$, используемый в $FCM-$алгоритме, а также значения $c_*$ и $c^*$, где $c_* \geq 2$ – наименьшее и $c^* \leq n-1$ – наибольшее из возможного числа классов в искомом разбиении, так что сущность $FCM-CV$-алгоритма заключается в построении множества нечетких $c$-разбиений для различных значений $c \in [c_*, c^*]$ и вычислении показателя оптимальности числа классов $V_c(P)$ для каждого $P$ с последующим выбором некоторого нечеткого разбиения $P*$, оптимального в смысле используемого показателя $V_c(P)$. Когда исследователь не обладает информацией, позволяющей определить значения $c_*$ и $c^*$, полагается $c_* := 2$ и $c^* := n-1$.

Вместе с тем при большом количестве $n$ объектов в исследуемой совокупности $X = \{x_1, ..., x_n\}$ число вычислений резко возрастает, поэтому возникает необходимость определения интервала $[c_*, c^*]$ значений наиболее возможного числа классов в искомом $P*$, для чего в работе [5] была предложена методика применения аппарата треугольных нечетких чисел. Сущность этой методики заключается в построении исследователем треугольного нечеткого числа $V = (\hat{c}, 2, n-1)$ с функцией представления формы $\mu_V(l)$ на основании задаваемого исследователем наиболее возможного $\hat{c}$ числа нечетких кластеров, представляющего собой модальное значение нечеткого числа $V$, и построением интервала достоверности $[c_*, c^*]$ как множества натуральных чисел, таких, что $c_* \geq c'$ и $c^* \leq c''$, где интервал $[c', c'']$ представляет собой $\alpha$-срез нечеткого числа $V$.

Таким образом, когда число классов в искомом нечетком $c$-разбиении $P*$ неизвестно и может быть определено лишь приблизительно, к примеру, на основании разведочного анализа данных [4], процесс классификации условно делится на два этапа: построение интервала достоверности $[c_*, c^*]$ числа классов в искомом $P*$ и обработка данных $FCM-CV$-алгоритмом в построенном интервале достоверности.

Вместе с тем *FCM* -алгоритм, отыскивающий минимум функционала (2), представляет собой параметрическое семейство по $\gamma$ при фиксированном числе кластеров $c$ [4], [7], так что при увеличении значения $\gamma$ возрастает неопределенность классификации, затрудняющая интерпретацию результатов, что выражается соотношением

$$\gamma \to \infty \Rightarrow \mu_{li} \to \frac{1}{c}, \forall l = 1,\ldots,c, \forall i = 1,\ldots,n, \tag{4}$$

что, в свою очередь, влияет на поведение показателей $V_c(P)$, поэтому при больших значениях $\gamma$ иногда оказывается невозможным определить локальный экстремум показателя $V_c(P)$ в интервале достоверности $[c_*, c^*]$. С этой целью вышеизложенная методология классификации может быть обобщена, для чего в $FCM - CV$ -алгоритме вместо величины $V_c(P)$ следует вычислять величину $V_c(P) \cdot \mu_V(l)$, где значения функции представления формы $\mu_V(l)$ определяются для всех $c \in [c_*, c^*]$.

В качестве примера для проведения вычислительного эксперимента были выбраны данные, использованные К.Г. Луни в качестве тестовых в работе [8]. Пятнадцать объектов исследуемой совокупности, структура которой изображена на рис. 1, представляют собой точки в двумерном признаковом пространстве.

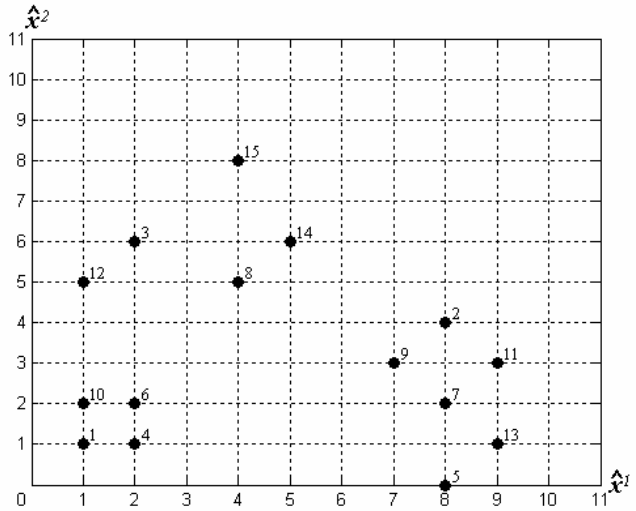


**Рис. 1. Исходные данные для вычислительного эксперимента**

Эксперименты проводились с использованием в качестве показателя $V_c(P)$ энтропии нечеткого $c$ -разбиения $V_{pe}(P)$, определяемой выражением [7]

$$V_{pe}(P) = -\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\left|\mu_{li} \cdot \ln \mu_{li}\right|, \tag{5}$$

где $n$ – число объектов исследуемой совокупности. При использовании $V_{pe}(P)$ условие (2) принимает вид

$$\min_c\left(V_{pe}(P)\right), c = 2,\ldots,n-1. \tag{6}$$

Результаты эксперимента, проведенного для $\hat{c} = 6, \alpha = 0{,}7$, представлены ниже. На рис. 2 изображено треугольное нечеткое число, соответствующее $\hat{c} = 6$, а также интервал достоверности $[c_* = 5, c^* = 8]$, соответствующий заданному значению порога $\alpha = 0{,}7$.

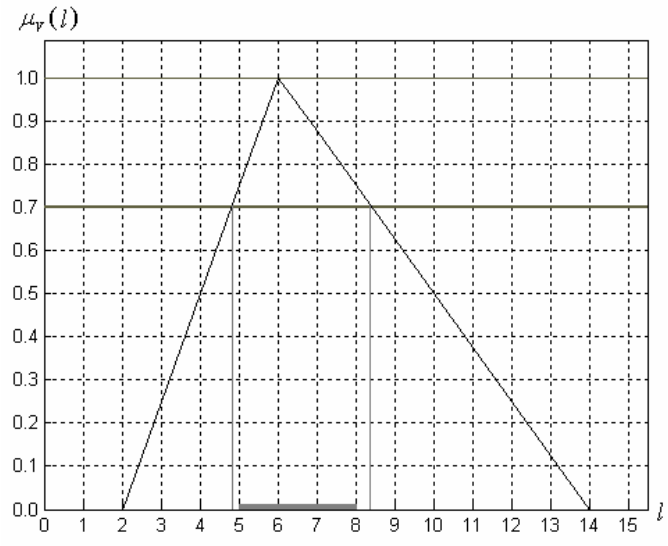


**Рис. 2. Функция представления формы треугольного нечеткого числа для $\hat{c} = 6$ и интервал достоверности $[c_* = 5, c^* = 8]$ при значении порога $\alpha = 0{,}7$**

На рис. 3 приведен график поведения величины $V_{pe}(P) \cdot \mu_V(l)$ в интервале $[c_* = 5, c^* = 8]$ при значении показателя нечеткости классификации $\gamma = 2$.

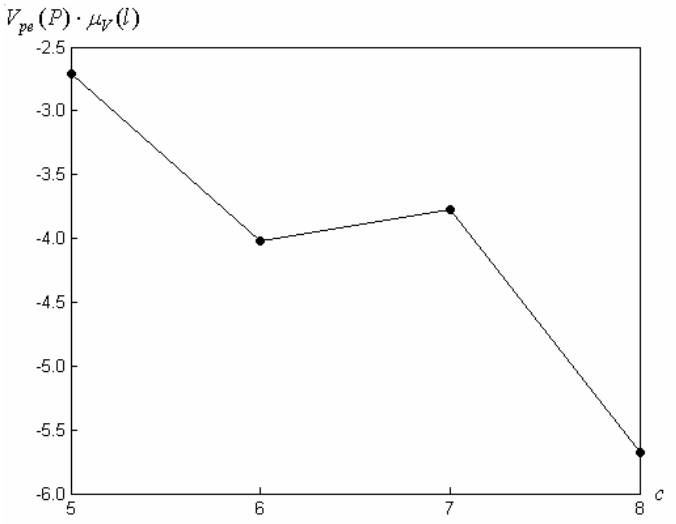


**Рис. 3. Значения величины $V_{pe}(P) \cdot \mu_V(l)$ нечеткого $c$-разбиения при $\gamma = 2$**

Результаты вычислительных экспериментов позволяют сделать вывод о высокой эффективности $FCM - CV$-алгоритма в сочетании с предлагаемым подходом к построению интервала достоверности $[c_*, c^*]$, основанного на применении нечетких чисел, благодаря значительному уменьшению вычислений при поиске нечеткого $c$-разбиения $P*$ на оптимальное число классов в полностью автоматическом режиме, а также возможности учета степени уверенности исследователя в принимаемом им решении о задаваемом в качестве параметра числе классов $c$, выражаемой в задаваемом значении порога $\alpha$, что является приемлемым с психологической точки зрения.

Поведение величины $V_{pe}(P) \cdot \mu_V(l)$ наглядно демонстрирует, что, как и при использовании энтропии $V_{pe}(P)$, исследуемая совокупность «расслаивается» на восемь

нечетких кластеров. Матрица соответствующего нечеткого $c$-разбиения $P*$ представлена в таблице.

Таблица 1

**Матрица нечеткого $c$-разбиения исследуемой совокупности на восемь классов**

| Номер объекта | Номер класса | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0,0054 | 0,7863 | 0,0072 | 0,0122 | 0,0046 | 0,1751 | 0,0048 | 0,0043 |
| 2 | 0,0061 | 0,0019 | 0,0026 | 0,0020 | 0,0059 | 0,0021 | 0,0247 | 0,9547 |
| 3 | 0,0104 | 0,0024 | 0,9421 | 0,0373 | 0,0009 | 0,0037 | 0,0014 | 0,0018 |
| 4 | 0,0042 | 0,8609 | 0,0050 | 0,0078 | 0,0041 | 0,1101 | 0,0043 | 0,0037 |
| 5 | 0,0046 | 0,0061 | 0,0033 | 0,0033 | 0,9200 | 0,0055 | 0,0421 | 0,0150 |
| 6 | 0,0081 | 0,1681 | 0,0106 | 0,0182 | 0,0052 | 0,7778 | 0,0062 | 0,0058 |
| 7 | 0,0020 | 0,0015 | 0,0012 | 0,0011 | 0,0184 | 0,0015 | 0,9572 | 0,0172 |
| 8 | 0,4878 | 0,0366 | 0,2030 | 0,1105 | 0,0206 | 0,0525 | 0,0357 | 0,0532 |
| 9 | 0,0409 | 0,0204 | 0,0202 | 0,0175 | 0,0667 | 0,0220 | 0,3632 | 0,4492 |
| 10 | 0,0051 | 0,1200 | 0,0077 | 0,0153 | 0,0031 | 0,8419 | 0,0035 | 0,0034 |
| 11 | 0,0234 | 0,0119 | 0,0123 | 0,0105 | 0,0775 | 0,0123 | 0,4402 | 0,4119 |
| 12 | 0,0005 | 0,0004 | 0,0023 | 0,9959 | 0,0001 | 0,0006 | 0,0001 | 0,0001 |
| 13 | 0,0122 | 0,0113 | 0,0079 | 0,0075 | 0,6614 | 0,0107 | 0,2290 | 0,0599 |
| 14 | 0,9658 | 0,0022 | 0,0123 | 0,0056 | 0,0018 | 0,0028 | 0,0034 | 0,0061 |
| 15 | 0,4149 | 0,0321 | 0,2865 | 0,1097 | 0,0233 | 0,0413 | 0,0366 | 0,0555 |

Вычисление величины $V_{pe}(P)\cdot\mu_V(l)$ вместо энтропии $V_{pe}(P)$ предпочтительно при достаточно больших значениях $\gamma$ и, кроме того, позволяет учитывать «степень достоверности» числа классов $c\in[c_*,c^*]$.

**Литература**


1. **Вятченин Д. А.** Нечеткие методы автоматической классификации. – Минск.: УП Технопринт, 2004. – 219 с.
2. **Вятченин Д. А., Садовская К. М.** Процедура поиска нечеткого разбиения множества объектов на оптимальное число классов//Сборник научных статей Военной академии Республики Беларусь. – 2004. – № 7. – С. 85–88.
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. Прикладная статистика: Классификация и снижение размерности: Справ. изд./ С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.
5. **Садовская К. М.** Применение нечетких чисел для построения интервала достоверности числа кластеров в нечетком разбиении//Вестник Военной академии Республики Беларусь. – 2005. – № 3. – С. 27–31.
6. **Соколов Б. В., Юсупов Р. М.** Концептуальные и методические основы квалиметрии моделей и полимодельных комплексов//Труды СПИИРАН. Вып. № 2 – СПб: СПИИРАН, 2004.
7. **Bezdek J. C.** Pattern Recognition with Fuzzy Objective Function Algorithms. – New York: Plenum Press, 1981. – 230 p.
8. **Looney C. G.** Interactive clustering and merging with a new fuzzy expected value // Pattern Recognition. – 2002. – Vol. 35. – P. 2413–2423.