# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

УДК 004.89: 004.021

С. И. Елизаров, М. С. Куприянов

## ПРОБЛЕМА ОПРЕДЕЛЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ ПРИ ИСПОЛЬЗОВАНИИ МЕТОДОВ РАЗБИЕНИЯ

Обсуждается проблема определения оптимального количества кластеров в методах разбиения. Рассматривается ряд известных критериев, используемых для центроидных методов: коэффициент разбиения, энтропия разбиения и эффективность разбиения. Предложены два новых критерия — для центроидного и нецентроидного методов: модифицированный коэффициент разбиения и качество разбиения.

***Ключевые слова:*** *кластеризация, методы разбиения, нечеткое отношение эквивалентности, критерий качества решения, коэффициент разбиения, энтропия разбиения.*

**Методы решения задачи кластеризации.** Задача кластеризации заключается в разбиении исследуемого множества объектов на группы, называемые кластерами, и решается в рамках первичного анализа данных. Среди ряда проблем, возникающих в ходе решения (выбор исходных признаков, выбор метода кластеризации, интерпретация результатов и др.), основной является проблема определения количества кластеров. Формально задача кластеризации описана в работе [1].

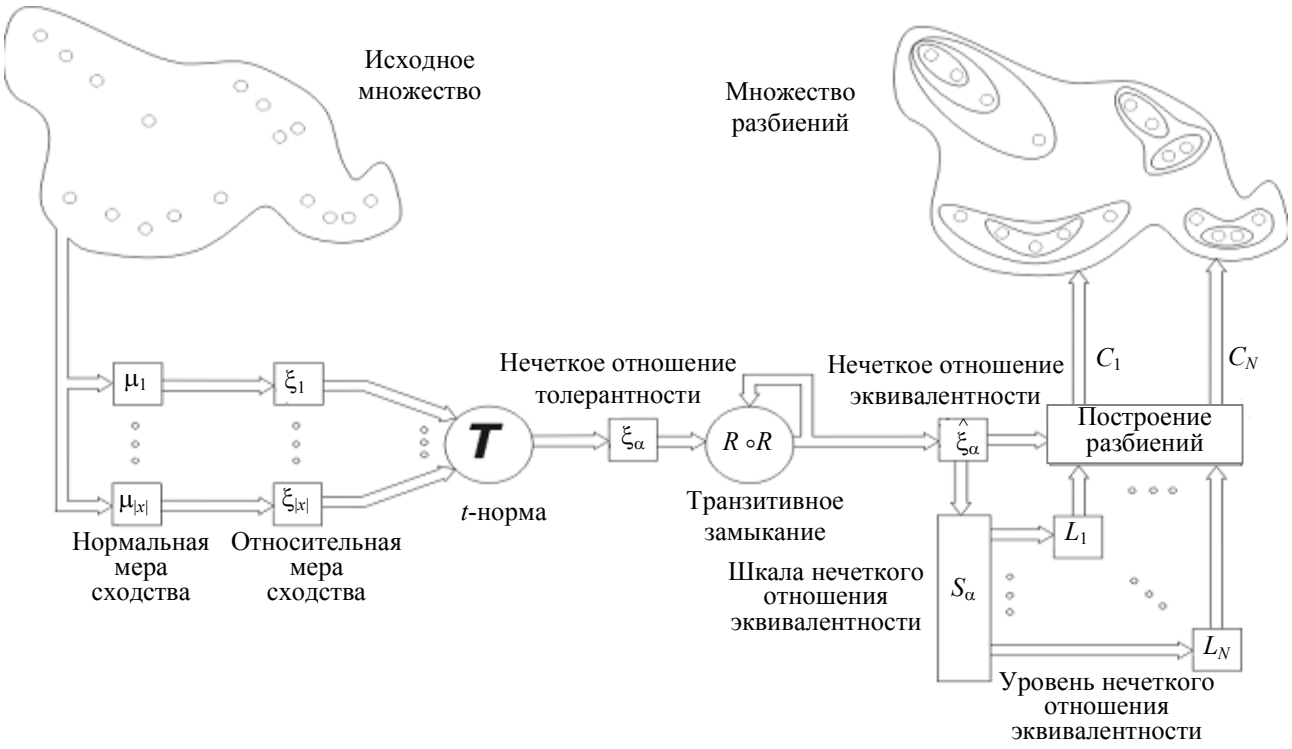
При решении задачи кластеризации наибольшее распространение получили методы построения разбиения, при использовании которых данные группируются в кластеры таким образом, чтобы целевая функция алгоритма разбиения достигала экстремума [2, 3]. Методы построения разбиения можно разделить на центроидные и нецентроидные. В центроидных методах вводится понятие центра кластера (центроида) — особой точки, вокруг которой формируется кластер. В большинстве практических приложений центроидные методы построения разбиения позволяют успешно решать задачу кластеризации. Однако эти методы имеют ряд общих недостатков, а именно:

— предполагается, что кластеры имеют форму эллипсоида: это далеко не всегда справедливо;

— в каждом кластере предполагается наличие некоторой узловой точки (центра кластера, центроида), степень принадлежности которой кластеру равна единице; степень принадлежности всех остальных точек данному кластеру меньше единицы и убывает по мере удаления точки от центра; при кластеризации данных со сложной внутренней структурой такое предположение о центре кластера не всегда приемлемо;

— построение кластеров на базе отношений между элементами входного множества и центрами кластеров, а не на базе отношений между самими элементами сужает круг возможных решений задачи кластеризации; существуют задачи, для которых верное решение при таком подходе вообще не может быть найдено.

Метод кластеризации на базе нечеткого отношения эквивалентности [4] предназначен для устранения недостатков, присущих центроидным методам. В данном методе на основе определения расстояния между каждой парой элементов входного множества строится сначала нечеткое отношение толерантности, затем при помощи транзитивного замыкания этого отношения получается нечеткое отношение эквивалентности, которое характеризуется шкалой уровней эквивалентности. Каждый из этих уровней соответствует некоторому отношению эквивалентности в классическом смысле. Схема кластеризации с использованием нечеткого отношения эквивалентности показана на рисунке [1].




К достоинствам метода можно отнести отказ от использования понятия центра кластера (решение задачи кластеризации находится исключительно при помощи отношений сходства между элементами входного множества), а также возможность поиска кластеров произвольной формы. Недостатком метода является чувствительность к шуму. Метод нечеткого отношения эквивалентности отличается от центроидных методов способом представления результатов (в виде матрицы нечеткого отношения эквивалентности).

Главный недостаток как центроидных, так и нецентроидных методов кластеризации заключается в необходимости указывать количество кластеров, на которое производится разбиение. Определение количества кластеров при решении задачи кластеризации обычно выводится за рамки задачи. Предполагается, что такая информация уже имеется к моменту начала ее решения. С другой стороны, даже при наличии некоторой информации о возможном количестве кластеров она, как правило, выражается не единственным числом, а интервалом. Для каждого допустимого количества кластеров находится соответствующее разбиение. Определить, какое разбиение является наиболее верным, довольно сложно. Отсюда можно сделать вывод, что выбор оптимального количества кластеров и выбор наилучшего решения задачи кластеризации из множества возможных — это две грани одной и той же проблемы. Чтобы решить эту проблему вводятся понятия критериев оценки качества решения [4, 5]. Использование этих критериев позволяет решить задачу определения количества кластеров и выбрать оптимальное решение.

**Критерии оценки качества решения задачи кластеризации.** Рассмотрим некоторые *критерии*, используемые *для центроидных методов построения разбиения.*

*Коэффициент разбиения* определяется как

$$K_{\text{р}} = \frac{\sum_{i=1}^{|X|}\sum_{j=1}^{|C|} u_{ij}^2}{|X|}, \qquad (1)$$

где $u_{ij}$ — элемент матрицы принадлежности, $X$ — входное множество, $C$ — множество кластеров.

Как видно из формулы (1), единственным аргументом данного критерия является матрица принадлежности.

Из определения функции принадлежности известно, что

$$u_{ij} \in [0,1], \ \sum_{j=1}^{|C|} u_{ij} = 1. \qquad (2)$$

Учитывая это, несложно показать, что данный критерий достигает минимума при всех $u_{ij} = |C|^{-1}$, $j = 1, \ldots, |C|$, и равен при этом $|C|^{-1}$. Это случай наибольшей неопределенности: все элементы входного множества с равной степенью принадлежат каждому из кластеров. Максимум данного критерия достигается на границе области определения ($u_{ip} = 1$, $u_{iq} = 0$, $p, q = 1, \ldots, |C|$, $p \neq q$), при этом критерий равен единице, что соответствует максимально четкому разбиению.

*Модифицированный коэффициент разбиения.* При применении коэффициента разбиения отмечено, что при малом количестве кластеров данный критерий не позволяет определить наилучшие разбиения. Очевидно, это связано с областью его значений. Не меняя характера критерия, изменим его таким образом, чтобы зависимость от количества кластеров не была связана с началом диапазона значений критерия. Для этого вычтем из коэффициента разбиения величину $|C|^{-1}$. В результате получим модифицированный коэффициент разбиения

$$K_{\text{р.м}} = \frac{\sum_{i=1}^{|X|}\sum_{j=1}^{|C|} u_{ij}^2}{|X|} - \frac{1}{|C|},$$

область значений которого будет находиться на отрезке $[0, (|C| - 1)/C]$. Применение модифицированного коэффициента разбиения дает более объективные результаты.

И коэффициент разбиения, и его модификации имеют неустранимый недостаток, связанный с оценкой разбиений при разном количестве кластеров, — области значений критериев напрямую зависят от выбранного количества кластеров, что ограничивает применение этих критериев.

Другой критерий — *энтропия разбиения* — определяется следующим выражением:

$$E_{\text{р}} = -\frac{\sum_{i=1}^{|X|}\sum_{j=1}^{|C|} u_{ij} \ln(u_{ij})}{|X|}.$$

Данный критерий строится по аналогии с определением, принятым в теории информации [6]. Определим, какие значения принимает этот критерий, учитывая выражение (2). Несложно показать, что он достигает максимума при всех $u_{ij} = |C|^{-1}$, $j = 1, \ldots, |C|$, и равен при этом $\ln |C|$; это соответствует наихудшему по информативности случаю: все значения функций принадлежности одинаковы. Минимальное значение данного критерия (равное нулю) будет получено в случае, если для каждого элемента входного множества найден кластер, принадлежность к которому равна единице, что соответствует четкому разбиению. Таким

образом, чем меньше значение данного критерия, тем более четкое разбиение будет получено. Тем не менее сравнивать при помощи данного критерия кластеризации, полученные при разном количестве кластеров, некорректно, поскольку диапазон его значений для каждой кластеризации будет разным.

*Модифицированная энтропия,* определяемая выражением

$$E_{\text{р.м}} = -\frac{\sum_{i=1}^{|X|}\sum_{j=1}^{|C|} u_{ij}\ln(u_{ij})}{|X|\ln|C|} = \frac{E_{\text{р}}}{\ln|C|},$$

лишена недостатков предыдущего критерия. Как видно, диапазон значений модифицированной энтропии не связан с количеством кластеров и лежит в отрезке [0,1]. В этом случае можно сравнивать результаты, полученные с использованием разного количества кластеров.

Следующий критерий — *эффективность разбиения* — определяется выражением

$$I_{\text{р}} = \sum_{j=1}^{|C|}\sum_{i=1}^{|X|} u_{ij}^2\left(d^2(k_j,\overline{x}) - d^2(x_i,k_j)\right),$$

где $\overline{x}$ — среднее значение элементов входного множества, $k_j$ — центр кластера $j$, $d(x,y)$ — расстояние между $x$ и $y$.

Данный критерий состоит из двух частей:

— межкластерные отличия

$$\sum_{j=1}^{|C|}\sum_{i=1}^{|X|} u_{ij}^2 d^2(k_j,\overline{x});$$

— внутрикластерные отличия

$$\sum_{j=1}^{|C|}\sum_{i=1}^{|X|} u_{ij}^2 d^2(x_i,k_j).$$

Получаем, что чем больше значения критерия, тем лучше выполнена кластеризация. Недостаток данного критерия заключается в том, что его значения выражены в абсолютных величинах (в единицах расстояния) и могут в общем случае быть произвольными, что усложняет интерпретацию: затруднительно сравнивать кластеризации, выполненные при различных наборах данных, даже если они взяты из одного множества.

Рассмотрим *критерий*, используемый *для метода нечеткого отношения эквивалентности*, — *качество разбиения*. При использовании нечеткого отношения эквивалентности для решения задачи кластеризации результатами являются матрица нечеткого отношения эквивалентности и шкала уровней эквивалентности. Каждый уровень шкалы порождает соответствующее разбиение на классы эквивалентности. Количество уровней в шкале эквивалентности велико и близко к мощности исследуемого множества. Значит, именно такое число различных решений можно получить, но не все они являются практически полезными.

При увеличении уровня эквивалентности новое разбиение на классы получается из предыдущего путем разделения одного или нескольких классов. Как правило, классы эквивалентности значительно различаются по мощности (особенно в первой половине шкалы отношения). Группу наиболее мощных классов назовем *практически полезными кластерами*. Данную группу необходимо уметь выделять из множества классов эквивалентности.

Пусть $C=\{c_1, \ldots, c_n\}$ — множество всех классов эквивалентности, полученных для данного уровня эквивалентности, $C_c \subseteq C$ — множество всех практически полезных кластеров. Необходимо определить минимальную мощность $N$ класса эквивалентности, при которой его

можно считать практически полезным кластером. В этом случае $C_c = \{\forall c_i \in C: |c_i| \geq N\}$. Для определения $N$ предлагается следующая процедура:

— классы эквивалентности распределяются по убыванию мощности;

— для каждой пары упорядоченных классов вычисляется взвешенное отношение

$$R = \frac{|c_i|}{|c_{i+1}|} \frac{|c_i| + |c_{i+1}|}{2},$$

где первый множитель — отношение двух соседних по мощности классов, а второй — весовой коэффициент, назначение которого (при $c_i = c_{i+1}$) — сместить максимум отношения $R$ ближе к началу последовательности;

— по максимуму $R$ определяется элемент в упорядоченной последовательности классов; его мощность будет равна $N$.

Реализуя описанную выше процедуру, для каждого разбиения на классы эквивалентности можно выделить группу практически полезных кластеров, что является важной составляющей решения задачи кластеризации. Другая числовая характеристика разбиения — коэффициент разбиения: отношение суммарной мощности кластеров к общей мощности множества.

Итак, для каждого результата кластеризации определены:

— уровень эквивалентности: чем он выше, тем более схожи элементы классов эквивалентности, тем более качественным будет разбиение;

— множество практически полезных кластеров: чем их больше, тем лучше;

— коэффициент разбиения: чем он больше, тем качественнее разбиение.

В общем случае по мере увеличения уровня эквивалентности множество практически полезных кластеров увеличивается, но уменьшается коэффициент разбиения. Исходя из этого можно сформировать критерий наилучшего разбиения: наилучшим назовем такое разбиение, для которого показатель качества разбиения

$$Q_i = L_i \cdot |C_c| \cdot \frac{\sum_{c_j \in C_c} |c_j|}{|X|}$$

достигает максимума; здесь $L_i$ — уровень эквивалентности, $C_c = \{c_j\}$.

Формула для определения показателя качества разбиения, как и процедура получения практически полезных кластеров выведена эмпирическим путем и на основании общих соображений о том, какое разбиение является полезным и наиболее качественным.

## СПИСОК ЛИТЕРАТУРЫ

1. Методы и модели анализа данных: OLAP и Data Mining / *А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод.* СПб.: БХВ-Петербург, 2004.

2. *Miyamoto S., Ichihashi H., Honda K.* Algorithms for Fuzzy Clustering: Methods in C-Means Clustering with Applications. Berlin: Springer, 2008.

3. *Jang J.-S.R., Sun C.-T., Mizutani E.* Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence. New Jersey: Prentice Hall, 1997.

4. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / *А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод.* СПб.: БХВ-Петербург, 2007.

5. *Зайченко Я. П.* Самообучение в интеллектуальных системах. Постановка задачи кластер-анализа. Критерии и метрики кластер-анализа [Электронный ресурс]: Основы проектирования интеллектуальных систем / *Я. П. Зайченко.* 2002: <http://www.iasa.org.ua/tpr.php?lang=rus&ch=2&sub=4 >.

6. *Стратонович Р. Л.* Теория информации. М.: Сов. радио, 1975.

***Сведения об авторах***

***Сергей Иванович Елизаров*** — аспирант; Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“, кафедра вычислительной техники; E-mail: duplex@rambler.ru

***Михаил Степанович Куприянов*** — д-р техн. наук, профессор; Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“, кафедра вычислительной техники; E-mail: Mikhail.kupriyanov@gmail.com

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 02.06.09 г.