Gorelova Galina Victorovna
Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University".
E-mail: g.v.gorelova@gmail.com.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634311426.
The Department of State and Municipal Legislation and Administration; Dr. of Eng. Sc.; Professor.

Drokina Christina Vladimirovna
E-mail: krdrokina@mail.ru.
Phones: +78634371704; +79185057127.
The Department of Management; Postgraduate Student; Assistant Lecturer.

УДК 519.23

А.В. Егоров, Н.И. Куприянова

ОСОБЕННОСТИ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДАННЫХ

Рассмотрены основные понятия кластеризации и нечеткой кластеризации данных. Описаны возможные типы данных, пригодных для кластеризации. Заданы исходные данные для алгоритмов кластеризации. Кратко проанализированы существующие алгоритмы кластеризации данных, отмечены их достоинства и недостатки. Описан наиболее перспективный нечеткий горный алгоритм кластеризации. Выявлены перспективы развития алгоритмов кластеризации как составной части математического аппарата поддержки интеллектуальных информационных систем.

Кластеризация; нечеткая кластеризация; горный алгоритм кластеризации.

A.V. Egorov, N.I. Kuprianova

CHARACTERISTICS OF METHODS OF FUZZY CLUSTERING DATA

The basic concepts of clustering and fuzzy clustering. The possible types of data suitable for clustering. Given input for clustering algorithms. Briefly analyzed the existing data clustering algorithms, their advantages and disadvantages. Described the most promising mining fuzzy clustering algorithm. Identified prospects of algorithms for clustering, as part of the mathematical tools to support intelligent information systems.

Clustering; fuzzy clusterin; mountain clustering algorithm.

Одним из направлений обработки данных различной структуры и свойств является кластеризация. Кластеризация – это объединение объектов в группы (кластеры) на основе схожести признаков для объектов одной группы и отличий между группами. Большинство алгоритмов кластеризации не опираются на традиционные для статистических методов допущения; они могут использоваться в условиях почти полного отсутствия информации о законах распределения данных [5]. Кластеризацию проводят для объектов с количественными (числовыми), качественными или смешанными признаками. Рассмотрим кластеризацию только для объектов с количественными признаками (отметив потенциал методов для качественных и смешанных признаков). Исходной информацией для кластеризации является матрица наблюдений:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}, \quad (1)$$

каждая строчка которой представляет собой значения n признаков одного из M объектов кластеризации. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров. В метрическом пространстве "схожесть" обычно определяют через расстояние. Расстояние может рассчитываться как между исходными объектами (строчками матрицы X), так и от этих объектов к прототипу кластеров. Обычно координаты прототипов заранее неизвестны – они находятся одновременно с разбиением данных на кластеры. Существует множество методов кластеризации, которые можно классифицировать на четкие и нечеткие [1]. Четкие методы кластеризации разбивают исходное множество объектов X на несколько непересекающихся подмножеств. При этом любой объект из X принадлежит только одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью. Методы кластеризации также классифицируются по тому, определено ли количество кластеров заранее или нет. В последнем случае количество кластеров определяется в ходе выполнения алгоритма на основе распределения исходных данных. Четкие методы являются наиболее изученными, рассмотрим некоторые из них.

Методы кластерного анализа можно разделить на две группы:

1. Иерархические методы (суть состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие).

1.1. Иерархические агломеративные методы (Agglomerative Nesting, AGNES). Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер. Минусом данного алгоритма является однофакторность процесса объединения кластеров с невозможностью учета группы схожих характеристик.

1.2. Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA). Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Недостатком иерархических алгоритмов является однофакторность процесса объединения кластеров с значительной сложностью учета группы схожих характеристик, а также невозможность реализовать данные методы на больших объемах данных.

2. Неиерархические (суть – в процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки) [5].

2.1. Алгоритм k-средних. Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Основными недостатками его являются возможное искажение среднего за счет выбросов и неэффективность работы при больших объемах данных. Несмотря на

это, стоит отметить, что данный алгоритм весьма распространен как нечеткий метод кластеризации количественный данных [3].

2.2. Алгоритм PAM (Partitioning Around Medoids). PAM является модификацией алгоритма k-средних, алгоритмом k-медианы (k-medoids).Алгоритм менее чувствителен к шумам и выбросам данных, чем алгоритм k-means, поскольку медиана меньше подвержена влияниям выбросов, эффективен для небольших объемов данных.

По итогам рассмотрения методов четкой кластеризации можно отметить:

1. Эти методы эффективны только на небольших объемах данных.

2. Результаты, полученные под действием данных методов, не позволяют продуктивно распределить элементы, находящиеся на границах кластеров.

3. Все методы начинаются с четкого задания количества кластеров, являющихся фиксированной величиной [2].

Таким образом, необходимо применять методы, учитывающие возможность работы с большим объемом данных качественного и количественного типа с использованием нечеткости и без предварительного задания кластеров. Наиболее подходящим при заданных условиях является модифицированный метод горной кластеризации. Метод предложен Р. Ягером и Д. Филевым в 1993 г. Кластеризация по горному методу не является нечеткой, однако ее часто используют при синтезе нечетких правил из данных. На первом шаге горной кластеризации определяют точки, которые могут быть центрами кластеров. На втором шаге для каждой такой точки рассчитывается значение потенциала, показывающего возможность формирования кластера в ее окрестности. Чем плотнее расположены объекты в окрестности потенциального центра кластера, тем выше значение его потенциала. После этого итерационно выбираются центры кластеров среди точек с максимальными потенциалами [5].

На первом шаге необходимо сформировать потенциальные центры кластеров. Для алгоритма горной кластеризации число потенциальных центров кластеров (Q) должно быть конечным. Ими могут быть объекты кластеризации (строчки матрицы X), тогда $Q = M$. Второй способ выбора потенциальных центров кластеров состоит в дискретизации пространства входных признаков. Для этого диапазоны изменения входных признаков разбивают на несколько интервалов. Проводя через точки разбиения прямые, параллельные координатным осям, получаем "решеточный" гиперкуб. Узлы этой решетки и будут соответствовать центрам потенциальных кластеров. Обозначим через q_r – количество значений, которые могут принимать центры кластеров по r-й координате ($r = \overline{1,n}$). Тогда количество возможных кластеров будет равно $Q = \prod_{r=\overline{1,n}} t_r$.

На втором шаге алгоритма рассчитывается потенциал центров кластеров по следующей формуле: $P(Z_h) = \sum_{k=\overline{1,M}}^{m} exp(-\alpha \cdot D(Z_h, X_k)),\ h = \overline{1,Q}$

где $Z_h = (z_{1,h}, z_{2,h}, ..., z_{n,h})$ – потенциальный центр h-го кластера;

α – положительная константа;

$D(Z_h, X_k)$ – расстояние между потенциальным центром кластера (Z_h) и объектом кластеризации (X_k). В евклидовом пространстве это расстояние рассчитывается по формуле $D(Z_h, X_k) = \sqrt{\|Z_h - X_k\|^2}$.

В случае, когда объекты кластеризации заданы двумя признаками (n=2), графическое изображение распределения потенциала будет представлять собой поверхность, напоминающую горный рельеф. Отсюда и название – горный метод кластеризации [1].

На третьем шаге алгоритма в качестве центров кластеров выбирают координаты "горных" вершин [5]. Для этого центром первого кластера назначают точку с наибольшим потенциалом. Обычно наивысшая вершина окружена несколькими достаточно высокими пиками. Поэтому назначение центром следующего кластера точки с максимальным потенциалом среди оставшихся вершин привело бы к выделению большого числа близко расположенных центров кластеров. Чтобы выбрать следующий центр кластера, необходимо вначале исключить влияние только что найденного кластера. Для этого значения потенциала для оставшихся возможных центров кластеров пересчитываются следующим образом: от текущих значений потенциала вычитают вклад центра только что найденного кластера (поэтому кластеризацию по этому методу иногда называют субтрактивной) [4]. Перерасчет потенциала происходит по формуле

$$P_2(Z_h) = P_1(Z_h) - P_1(V_1) \cdot exp(-\beta \cdot D(Z_h, V_1)),$$

где $P_1(.)$ – потенциал на 1-й итерации, $P_2(.)$ – потенциал на 2-й итерации, V_1 – центр первого найденного кластера:

$$V_1 = \underset{Z_1, Z_2, \ldots, Z_Q}{\arg} \max(P_1(Z_1), P_1(Z_2), \ldots, P_1(Z_Q));$$

β – положительная константа.

Центр второго кластера определяется по максимальному значению обновленного потенциала:

$$V_2 = \underset{Z_1, Z_2, \ldots, Z_Q}{\arg} \max(P_1(Z_1), P_1(Z_2), \ldots, P_1(Z_Q)).$$

Затем снова пересчитывается значение потенциалов:

$$P_3(Z_h) = P_2(Z_h) - P_2(V_2) \cdot \exp(-\beta \cdot D(Z_h, V_2)).$$

Итерационная процедура пересчета потенциалов и выделения центров кластеров продолжается до тех пор, пока максимальное значение потенциала превышает некоторый порог.

На основе проведенного анализа можно сказать, что на сегодняшний день различные алгоритмы кластеризации в зависимости от типа исходных данных и точности распределения нашли широкое распространение. Наиболее перспективными являются нечеткие алгоритмы, так как они обрабатывают значительные объемы данных и позволяют распределить элементы, находящиеся на границе кластеров. В основном все рассмотренные методы оперируют с количественными данными, что позволяет расширить их алгоритмы также для смешанных и качественных признаков. Пошаговость дает возможность программной реализации алгоритмов, что существенно сократит время расчетов и увеличит результативность процедуры кластеризации.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Babuska R.* Fuzzy Modeling for Control.-Boston: Kluwer Academic Publishers, 1998.
2. *Gustafson D.E., Kessel W.C.* Fuzzy Clustering with a Fuzzy Covariance Matrix. San-Diego, USA, 1979. – P. 761-766.
3. *Xei X.L., Beni G.A.* Validity Measure for Fuzzy Clustering // IEEE Transaction on Pattern Analysis and Machine Intelligent. – 1991. – 3 (8). – P. 841-846.
4. *Yager R., Filev D.* Essentials of Fuzzy Modeling and Control. USA: John Wiley & Sons, 1984. – P. 387.
5. *Чубукова И.А.* Data Mining. – М.: Интернет-Университет информационных технологий; БИНОМ. Лаборатория знаний, 2006.

Статью рекомендовал к опубликованию д.т.н., профессор В.П. Карелин.

Егоров Александр Вадимович
Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: egor@tsure.ru.
347928, г. Таганрог, пер. Некрасовский, 44.
Тел.: 88634383652.
Кафедра прикладной информатики; к.т.н.; доцент.

Куприянова Наталия Игоревна
E-mail: ultra-n@list.ru.
Кафедра прикладной информатики; аспирант.

Yegorov Alexander Vadimovich
Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: egor@tsure.ru.
44, Nekrasovskiy, Taganrog, 347928, Russia.
Phone: +78634383652.
The Department of Applied Information Science; Cand. of Eng. Sc.; Associate Professor.

Kupriyanova Natalia Igorevna
E-mail: ultra-n@list.ru.
The Department of Applied Information Science; Postgraduate Student.

УДК 338.48

Н.А. Карастелкина

КОГНИТИВНЫЙ ПОДХОД К ПОСТРОЕНИЮ ТУРИСТСКО-РЕКРЕАЦИОННОГО КЛАСТЕРА В РЕГИОНЕ

Уровень развития туризма на современном этапе требует применения новых современных технологий изучения и развития туристской сферы. Наиболее действенным инструментом, который может позволить системно изучить туристскую отрасль, является когнитивный подход. Приведение же самой туристской сферы к такой системе, как туристский кластер позволяет повысить эффективность результатов туристской деятельности, поскольку грамотная организация четко структурированных составляющих системы может принести эффективную отдачу от вложенных усилий.

Когнитивный подход; рекреационная система; кластер.

N.A. Karastelkina

COGNITIVE APPROACH TO CONSTRUCTION OF TURIST-RECREATIONAL CLUSTER IN REGION

At the present stage the level of development of tourism demands application of new modern technologies of studying and development of tourist sphere. The most effective tool which can presumes to study tourist branch as a system is a cognitive approach. Reduction of tourist sphere to such system as tourist cluster allows to raise efficiency of results of tourist activity because of the competent organization of accurately structured components of system can bring effective return from the enclosed efforts.

The cognitive approach; recreational system; cluster.