УДК 519.23

Н.И. Куприянова

КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ КЛАСТЕРИЗАЦИИ ДАННЫХ

Сформирована и проанализирована концептуальная модель кластеризации данных. Заданы начальные характеристики данной модели, описываются возможные алгоритмы, реализуемые в ее рамках. Каждый алгоритм оценивается с точки зрения применяемости к различным типам данных и описывается с точки зрения его параметров и универсальности. Особое место уделяется анализу группы нечетких алгоритмов. Параллельно формулируются понятия «нечеткая и лингвистическая переменная». Каждый нечеткий алгоритм задается на основе нечеткой и лингвистической переменной. В статье также освещаются функции принадлежности для ввода и вывода данных в концептуальной модели нечеткой кластеризации.

Концептуальная модель кластеризации; нечеткая переменная; лингвистическая переменная; алгоритм кластеризации.

N.I. Kupriyanova

CONCEPTUAL MODEL OF CLUSTERING DATA

This article formed the conceptual model and analyzed the clustering of data. Given the initial characteristics of this model describes the possible algorithms are implemented within its framework. Each algorithm is evaluated in terms of applicability to different types of data and is described in terms of options and versatility. Particular attention is given to the analysis of fuzzy algorithms. In parallel, formulated the concept of "fuzzy and linguistic variables. Each fuzzy set algorithm based on fuzzy and linguistic variables. The article also highlights the membership functions for input and output data in a conceptual model of fuzzy clustering.

Conceptual model of clustering; fuzzy variable; linguistic variable; the clustering algorithm.

Под концептуальной моделью понимается модель предметной области, состоящей из перечня взаимосвязанных понятий, используемых для описания этой области, вместе со свойствами и характеристиками, классификацией этих понятий, по типам, ситуациям, признакам в данной области и законов протекания процессов в ней [1].

Рассмотрим на основе данного определения процесс кластеризации данных.

Он может реализовываться как кластеризация множества характеристик одного объекта или множества объектов при делении их на кластеры. В первом случае модель начинает описываться с помощью формальных определений.

Они описывают процесс кластеризации с точки зрения активно используемых структур data mining. Вектор характеристик (объект) x – единица данных для алгоритма кластеризации. Обычно это элемент d -мерного пространства: $x = (x_1,..., x_d)$. Характеристика (атрибут) x_i – скалярная компонента вектора x. Размерность d – количество характеристик объекта x. Множество объектов $x = (x_1,..., x_n)$ – набор входных данных. i-й объект из X определяется как $x_i = (x_{1,i},..., x_{i,d})$. Часто X представляют в виде матрицы характеристик размера n x d. Кластер – подмножество «близких друг к другу» объектов из X. Расстояние $d(x_i, x_j)$ между объектами x_i и x_j – результат применения выбранной метрики (или квазиметрики) в пространстве характеристик. Данные определения – начальные как в формировании кластеров и кластеризации, так и начальные в множестве задач data mining [2].

Под кластерным анализом для множества объектов понимается задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. С точки зрения концептуальной модели опишем процесс кластеризации формально: Пусть X – множество объектов, Y(формально) — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, ..., x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера. Номер кластера может задаваться явно – как конкретное значение y_i, так и функция принадлежности (Membership Function) через $MF_c(x)$ – степень принадлежности к нечеткому множеству C, представляющей собой обобщение понятия характеристической функции обычного множества [2].

Таким образом, явно описываются возможности процесса кластеризации данных как:

- Простая кластеризация.
- Нечеткая кластеризация данных.
- Кластеризация нечетких данных.
- Нечеткая кластеризация нечетких данных.

Данные альтернативы реализуются с помощью алгоритмов кластеризации.

Алгоритм кластеризации – это функция, которая любому объекту ставит в соответствие номер кластера. Множество в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации.

Рассмотрим стандартные алгоритмы кластеризации, а затем опишем возможную альтернативу применениях их к нечетким данным или преобразуем в нечеткие алгоритмы:

1. Алгоритмы иерархической кластеризации. Среди алгоритмов иерархической кластеризации выделяются два основных типа: восходящие и нисходящие алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: вначале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Для вычисления расстояний между кластерами чаще все пользуются двумя расстояниями: одиночной связью или полной связью. К недостатку иерархических алгоритмов можно отнести систему полных разбиений, которая может являться излишней в контексте решаемой задачи.

2. Алгоритмы квадратичной ошибки. Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения. Самым распространенным алгоритмом этой категории является метод k-средних.

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

3. Алгоритмы, основанные на теории графов. Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа G=(V, E), вершинам которого соответствуют объекты, а ребра имеют вес, равный «расстоянию» между объектами. Достоинством графовых алгоритмов кластеризации являются наглядность, относительная простота реализации и возможность внесения различных усовершенствований, основанные на геометрических соображениях. Основным алгоритмом является алгоритм послойной кластеризации.

4. Алгоритм выделения связных компонент. В алгоритме выделения связных компонент задается входной параметр R и в графе удаляются все ребра, для которых «расстояния» меньше R. Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение R, лежащее в диапазоне всех «расстояний», при котором граф «развалится» на несколько связных компонент. Полученные компоненты и есть кластеры. Параметр R подбирается из зоны минимума между этими пиками. При этом управлять количеством кластеров при помощи порога расстояния довольно затруднительно.

5. Алгоритм минимального покрывающего дерева. Алгоритм минимального покрывающего дерева сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом.

6. Послойная кластеризация. Алгоритм послойной кластеризации основан на выделении связных компонент графа на некотором уровне расстояний между объектами (вершинами). Алгоритм послойной кластеризации формирует последовательность подграфов графа G, которые отражают иерархические связи между кластерами. Посредством изменения порогов расстояния возможно контролировать глубину иерархии получаемых кластеров. Таким образом, алгоритм послойной кластеризации способен создавать как плоское разбиение данных, так и иерархическое.

Анализ на основе сравнения для алгоритмов кластеризации представлен в табл. 1.

Таблица 1

Алгоритм кластеризации	Форма кластеров	Входные данные	Результаты
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
k-средних	Гиперсфера	Число кластеров	Центры кластеров
c –средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Выделение связных компонент	Произвольная	Порог расстояния R	Древовидная структура кластеров
Минимальное покрывающее дерево	Произвольная	Число кластеров или порог расстояния для удаления ребер	Древовидная структура кластеров
Послойная кластеризация	Произвольная	Последовательность порогов расстояния	Древовидная структура кластеров с уровнями иерархии

Далее рассмотрим нечеткую кластеризацию.

В большинстве современных работ данный тип кластеризации реализуется с помощью степени принадлежности к определенному кластеру С ($MF_c(x)$). Тогда нечетким кластером С называется множество упорядоченных пар вида $C = \{MF_C(x)\} MF_C(x) \in [0,1]$. Значение $MF_C(x)=0$ означает отсутствие принадлежности к множеству, 1 – полную принадлежность.

Наиболее популярными являются алгоритм нечеткой самоорганизации c-means и его обобщение в виде алгоритма Густафсона-Кесселя. Это стандартные четкие иерархические алгоритмы, использующие меру принадлежности к различным кластерам. Используя матрицу принадлежности, ищется значение критерия нечеткой ошибки и затем реализуется ее уменьшение. Формирование такого алгоритма близко к возможностям метода квадратичной ошибки в четкой кластеризации. Если реберные веса представить в интервале от 0 до 1 – метод минимального покрывающего дерева и алгоритм связанных компонент как наглядно, так и функционально реализует возможности нечеткости и данных функциональных механизмов data mining.

Адаптивно возможно и использование популярного алгоритма горной кластеризации. Он близок к c-means алгоритму, но каждый кластер, характеризуясь 2-мя координатами, дает толчок к формированию нечетких правил, на основе которых и строится нечеткая принадлежность в разрезе разделения всех элементов на кластеры.

Таким образом, получаем тенденцию к интерпретации алгоритмов в разрезе концептуальной модели за счет описания в нечеткости матрицы принадлежности.

Также алгоритм кластеризации может быть реализован на изначально нечетких и лингвистической переменных.

Нечеткая переменная описывается набором (N,X,A), где N – это название переменной, X – универсальное множество (область рассуждений), A – нечеткое множество на X [3].

Значениями лингвистической переменной могут быть нечеткие переменные, т.е. лингвистическая переменная находится на более высоком уровне, чем нечеткая переменная. Каждая лингвистическая переменная состоит:

- из названия;
- множества своих значений, которое также называется базовым терм-множеством T.

Элементы базового терм-множества представляют собой названия нечетких переменных:

- универсального множества X;
- синтаксического правила G, по которому генерируются новые термы с применением слов естественного или формального языка;
- семантического правила P, которое каждому значению лингвистической переменной ставит в соответствие нечеткое подмножество множества X.

Существует свыше десятка типовых форм кривых для задания функций принадлежности. Наибольшее распространение получили: треугольная, трапецеидальная и гауссова функции принадлежности [4].

Треугольная функция принадлежности определяется тройкой чисел (a,b,c), и ее значение в точке x вычисляется согласно выражению

$$MF(x)=\begin{cases}1-\dfrac{b-x}{b-a}, a\le x\le b,\\ 1-\dfrac{x-b}{c-b}, b\le x\le c,\\ 0, \textit{в остальных случаях}.\end{cases} \quad (1)$$

При (b–a) = (c–b) имеем случай симметричной треугольной функции принадлежности, которая может быть однозначно задана двумя параметрами из тройки (a,b,c).

Аналогично для задания трапецеидальной функции принадлежности необходима четверка чисел (a,b,c,d), описанная формулой

$$MF(x)=\begin{cases}1-\dfrac{b-x}{b-a}, a\le x\le b,\\ 1, b\le x\le c,\\ 1-\dfrac{x-c}{d-c}, c\le x\le d,\\ 0, \textit{в остальных случаях}.\end{cases} \quad (2)$$

При (b–a) = (d–c) трапецеидальная функция принадлежности принимает симметричный вид.

Функция принадлежности гауссова типа описывается формулой

$$MF(x)=\exp\left[-\left(\frac{x-c}{\sigma}\right)^2\right]. \quad (3)$$

Зачастую треугольную и трапецеидальную функцию используют для ввода и вывода в кластеризуемой системе. В отечественных и западных источниках возможности кластеризации нечетких данных реализованы сжато, без подробного описания, и именно в этой области кластеризации существует свободная ниша для дальнейших исследований.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Толковый словарь по искусственному интеллекту / *Авторы-составители А.Н. Аверкин, М.Г. Гаазе-Рапопорт, Д.А. Поспелов.* – М.: Радио и связь, 1992. – 256 с.
2. *Чубукова И.А.* Data Mining: Учебное пособие. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.
3. *Дюк В., Самойленко А.* Data Mining. – СПб.: Питер, 2001. – 368 с.
4. *Журавлев Ю.И., Рязанов В.В., Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.

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

Куприянова Наталия Игоревна – Технологический институт федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге; e-mail: ultra-n@list.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371743; кафедра прикладной информатики; аспирантка.

Kupriyanova Natalia Igorevna – Taganrog Institute of Technology – Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education “Southern Federal University”; e-mail: ultra-n@list.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; тел.: +78634371743; the department of applied information science; postgraduate student.