УДК 004.8

ИНТЕРАКТИВНАЯ КЛАСТЕРИЗАЦИЯ НА БАЗЕ НЕЧЕТКИХ ОТНОШЕНИЙ И ТЕХНОЛОГИИ VISUAL MINING

Виноградов Г.П., Мальков А.А.
(Тверской государственный технический университет, Тверь)
kja227@list.ru

Рассматриваются и анализируются методы кластеризации, основанные на четком и нечетком отношении сходства элементов выборки из генеральной совокупности. Предлагается подход к решению проблемы на базе интерпретации результатов, полученных после кластеризации, основанный на технологии Visual Mining.

Ключевые слова: кластер, Visual Mining, интерактивная оптимизация.

Введение

В последние годы в связи с появлением технологии Data Mining интерес к проблеме разбиения множества данных на кластеры усилился, т.к. применение соответствующих алгоритмов позволяет перейти от описания сравнительно небольших выборок объектов к описанию классов объектов и отношений между ними в хранилищах данных, измеряемых терабайтами. Однако, алгоритмов кластеризации с приемлемым уровнем качества, работающих с различными типами данных, со сверхбольшими объемами данных, в условиях наличия различного рода помех, ускорения темпов изменения состояния объектов нет, что отмечалось рядом исследователей [1]. Наиболее распространенные алгоритмы кластеризации (к ним относятся различ-

ные модификации алгоритмов k-means, Expectation Maximization и др.) в явной или неявной форме используют предположение, что каждый кластер представляет собой выпуклое множество. В ряде практических задач такое предположение не выполняется и приводит к неверным результатам.

1. Теоретический анализ

Пусть производятся наблюдения над объектами $x \in X$ из генеральной совокупности X. Пусть каждый объект появляется случайно и независимо от другого. Пусть выполняется «гипотеза разделимости», которая предполагает существование в анализируемых данных структуры. Это означает, что в анализируемых данных существует m классов и каждый объект может быть отнесен к одному из них, т.е. $x_i \in \Omega_k$, $k=1,...,m$, $i=1,...,N$, где m – число кластеров, N – число наблюдаемых объектов.

Если пространство X является непрерывном пространством, в котором согласно «гипотезе разделимости» существуют определенные связи (общие «свойства») между векторами $x_i \in X$, $i=1,...,N$ (например, метрические, вызывающие появление областей локальных сгущений объектов), то пространство

$$(1) \quad \Omega = \bigcup_{i=1}^{m} \Omega_k$$

является дискретным пространством. Очевидно, что пространство Ω накрывает пространство X и получается путем применения нелинейного правила преобразования G к пространству X: $G: X \to A$.

Задача кластеризации состоит в определении номера класса k для n-мерного вектора $x_i=\{x_{ij}, j=1,...,n\}$ из заранее неизвестного количества классов. То есть разбиение данных по кластерам (области сгущения точек) выполняется при одновременном определении их количества и геометрических характеристик.

Методы кластерного анализа для заданного набора данных, для которых каждая характеристика задается четким числом, а класс принадлежности каждого объекта неизвестен и количество

классов неизвестно, характеризуются следующими параметрами:

используемым способом сравнения данных между собой (мерой сходства);

методом (правилом) объединения точек в кластеры.

Пусть $d(x_i,x_j)$ – мера сходства объектов, тогда кластер Ω_k будет формироваться, например, по правилу:

(2) $\Omega_k = \{x_i,x_j \mid x_i \in X,\ x_j \in I \text{ и } d(x_i,x_j) \le \sigma\}$,

где σ – величина, определяющая меру сходства (близости) для включения объектов в один кластер.

Для меры близости $d(x_i,x_j)$ должны выполняться следующие условия:

$d(x_i,x_j) = d(x_j,x_i)$;

$d(x_i,x_j) = 0$ тогда и только тогда, когда $x_i = x_j$;

$d(x_i,x_j) \le d(x_i,x_p) + d(x_p,x_j)$.

Правило (2), очевидно, определяет дискретную характеристическую функцию, принимающую одно из двух значений: принадлежит или не принадлежит объект данных заданному классу.

Алгоритмы, использующие правило (2), образуют класс алгоритмов четкой кластеризации, которые в той или иной форме накладывают ограничения на геометрию получаемых кластеров, требуя охвата каждого кластера отдельным выпуклым множеством.

Известные методы в отличие от предлагаемого используют различные метрические меры для определения сходства/различия между объектами.

Поскольку кластер Ω_k является подмножеством множества X, т.е. $\Omega_k \subseteq X$, то для каждой его точки можно указать меру степени принадлежности к кластеру $\mu_{\Omega k}(x_i) \in [1 \div 0]$. Чем ближе значение функции к единице, тем больше степень уверенности в том, что данная точка принадлежит к кластеру с номером k.

Второй подход основан на введении меры сходства j-го элемента последовательности с i-м элементом на основе метрических мер, например, по правилу [2]:

$$(3)\quad \mu_i(j) = 1 - \frac{d(x_i, x_j)}{max\{d(x_i, x_k),\ k \in [\overline{1,n}]\}},\quad i,j \in [\overline{1,n}]$$

В первом случае исходное невыпуклое множество аппроксимируется набором выпуклых множеств. На основе полученных результатов строится логическая функция – высказывание о принадлежности элемента тому или иному кластеру.

Второй подход позволяет получать кластеры произвольной формы, но при наличии у кластеров близких точек или пересечений требует принятия дополнительных условий для определения номера кластера. Это, в свою очередь, при интерпретации полученных результатов предполагает выполнение значительной аналитической работы экспертом-аналитиком, связанной с просмотром большого числа вариантов.

Получающиеся при этом алгоритмы достаточно просты, но положенные в их основу эвристики на одних и тех же данных порождают, вообще говоря, различные классификации. Кроме того, результат их работы может и не иметь достаточного статистического обоснования.

2. Описание алгоритма

Использование функций принадлежности позволяет вводить в оценку меры сходства/различия объектов субъективные оценки эксперта-аналитика, отражающие его опыт, знание об исследуемом явлении, квалификацию, предпочтения и т.п.

Проблемой здесь является способ получения подобной информации. Одним из средств решения этой проблемы является технология визуального анализа данных – Visual Mining, которая позволяет человеку работать с визуальным представлением данных и результатами их обработки, понять их суть, напрямую взаимодействовать с данными и делать выводы путем использования различного вида графических образов, удобных для восприятия человеком [2]. Использование технологии Visual Mining позволяет представить процесс поиска «естественных» кластеров как процесс генерации гипотез экспертом с последующей их

проверкой алгоритмическими средствами, например, описанными выше с соответствующей доработкой. Примерную структурную схему гибридной визуальной кластеризации можно представить рис.1.

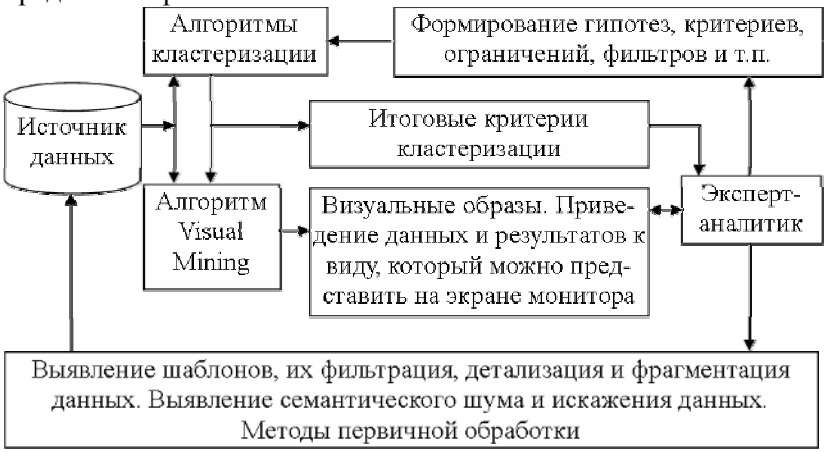

Рис.1. Схема кластеризации на основе технологии Visual Mining

Такая комбинация технологий позволит:

- мягко работать с неоднородными и зашумленными данными, т.к. эксперт-аналитик при этом использует «контекстную информацию» об изученном процессе, а алгоритмические методы не могут ее использовать;
- получить высокую степень конфиденциальности полученных результатов, т.к. соответствующая «теория», объясняющая структуру данных находится в голове аналитика;
- ускорить процесс и качество полученных знаний.

Здесь данные извлекаются из некоторого источника данных. Это может быть хранилище данных, ряд баз данных и т.п. К ним применяются методы кластерного анализа, его результаты и исходные данные подвергаются обработке методами Visual Mining для получения графических образов, пригодных для визуального восприятия пользователем. Пользователь на основе их восприятия формирует гипотезу о структуре кластеров, определяет меры сходства объектов, ограничения для алгорит-

мов и т.д. Одновременно, зная контекст для полученных данных, он идентифицирует шаблоны данных, выделяет данные к ним относящиеся для более детального анализа, выявляет факты искажения данных. Описанная схема позволяет эксперту «погрузиться» в данные, работать с их визуальным представлением, напрямую взаимодействовать с ними и получить наиболее адекватную их модель (извлечь знания из данных).

Пусть множество A вариантов допустимых разбиений на кластеры сформировано и каждый из них отличается от другого видом функции принадлежности. Обозначим через z_j, $j=1,\ldots,m$ критерии, которыми пользуется эксперт-аналитик при оценке качества разбиения. Тогда при наличии у него достаточной квалификации можно предположить наличие образа Y множества A с помощью отображения $Z:A\rightarrow R$, причем $Y=Im\ Z$. Следовательно, Y – это подмножество R^m и является множеством допустимых векторных оценок. Очевидно, что можно указать эффективную границу этого множества, если эксперт может установить естественный полупорядок на множестве R^m:

(4) $$eff(Y)=\{y_2\in Y | y_1\succ y_2 \& y_1\in Y \Rightarrow y_1=y_2\}.$$

В этом случае возможно применение рационального принципа принятия решения S по правилу:

$$S(A,Z)\subset A;$$

$$z(a)=z(b), \text{ если } a\in S(A,Z)\subset A \text{ и } b\in S(A,Z)\subset A;$$

$$S(A,Z)\subset z^{-1}(eff(Y));$$

$$S(A,Z)\neq\varnothing.$$

Первое условие определяет, что решениями могут быть только допустимые действия. Второе – два одинаковых действия, имеющие одинаковые векторные оценки, должны быть решениями. Третье условие определяет выбор только на множестве эффективных оценок. Такие оценки называются оптимальными по Парето. Определение вариантов разбиения согласно принципу Парето представляет собой классическую задачу векторной оптимизации:

(5) $\{a \in A | z(a) \in eff(Y)\}$.

И последнее условие означает, что аналитик способен сгенерировать по крайней мере одно допустимое решение.

Сформулированные принципы принятия рационального выбора являются менее жесткими, чем, например, условия существования совершенных отношений предпочтения. Это связано с тем, что ранжированию подлежат элементы множества *z(a)*, а не сами действия *A*, что существенно уменьшает нагрузку на аналитика. Это позволило для практической реализации принципа рационального принятия решения использовать схему, описанную в [3].

3. Экспериментальная часть

Работа описанного выше метода проверялась на ряде тестовых примеров с использованием алгоритма самоорганизации для формирования множества возможных вариантов кластеризации.

Такого рода алгоритм начинает работу с определения количества кластеров – находится первоначальное расположение центров каждого кластера. На этом этапе может возникнуть проблема: найденное количество кластеров окажется неверным. Визуализация полученных результатов позволяет эксперту-аналитику быстро определить примерное количество кластеров, распределение центров и рассчитать примерное положение центров кластеров, например, с помощью алгоритма разностного группирования. Когда будет найдено удовлетворительное первоначальное распределение центров кластеров, запускается алгоритм уточнения положения этих центров. При этом каждый вектор выборки будет принадлежать конкретному кластеру в некоторой степени. Алгоритм заканчивает работу, когда элементы матрицы степеней принадлежности будут различаться не более чем на заданную величину на двух ближайших итерациях.

В таблице 1 приведены центры и разбросы для распределения точек для примеров 1, 2. Координаты точек генерировались случайным образом.

Таблица 1

Координаты центра (x;y)	Номер примера			
	1		2	
	Разброс по x	Разброс по y	Разброс по x	Разброс по y
(0;5)	± 1.5	± 0.2	± 1.3	± 0.2
(0;-5)	± 1.5	± 0.2	± 1.3	± 0.2
(0.7;0)	± 0.6	± 4	± 0.8	± 4
(-0.7;0)	± 0.6	± 4	± 0.8	± 4

На рис.2, 3 изображены «истинные» центры кластеров (изображены прямоугольниками), траектории уточнения положений центров кластеров, а также в виде графиков представлены найденные алгоритмом значения степеней принадлежности каждой точки тому или иному кластеру для примеров 1 и 2 соответственно. Как показали вычисления, алгоритм самоорганизации нечувствителен к начальному расположению центров кластеров. Для этого алгоритма важно только правильное количество кластеров. Для учета формы разброса точек кластеров в алгоритме использовались масштабирующие и ковариационные матрицы, позволяющие с большой точностью определить форму кластеров и их расположение в пространстве.

В таблице 2 приведены координаты найденных алгоритмом самоорганизации центров для каждого примера и суммы квадратов отклонений найденных от «истинных» центров.

Таблица 2

Пример 1	Пример 2
(0.0967; 4.9336)	(0.0177; 4.7212)
(-0.0307; -4.8327)	(0.0763; -4.6861)

(0.7116; 0.0631)	(0.2607; -1.3081)
(-0.6966; 0.0787)	(-0.3662; 1.4117)
Суммы квадратов отклонений	
0.0530	4.1907

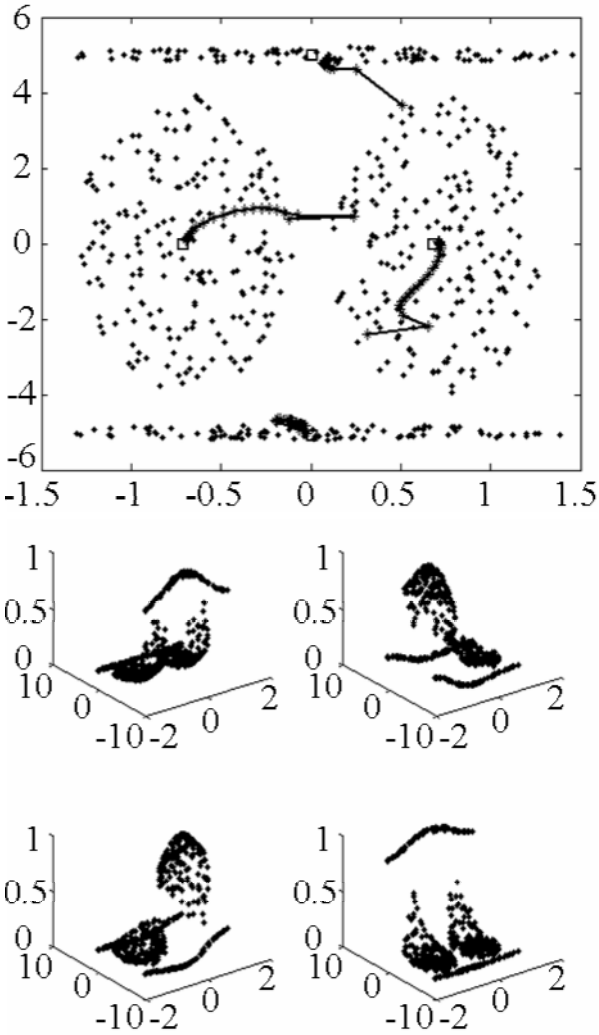

Рис.2. Результат определения параметров кластеров для примера 1

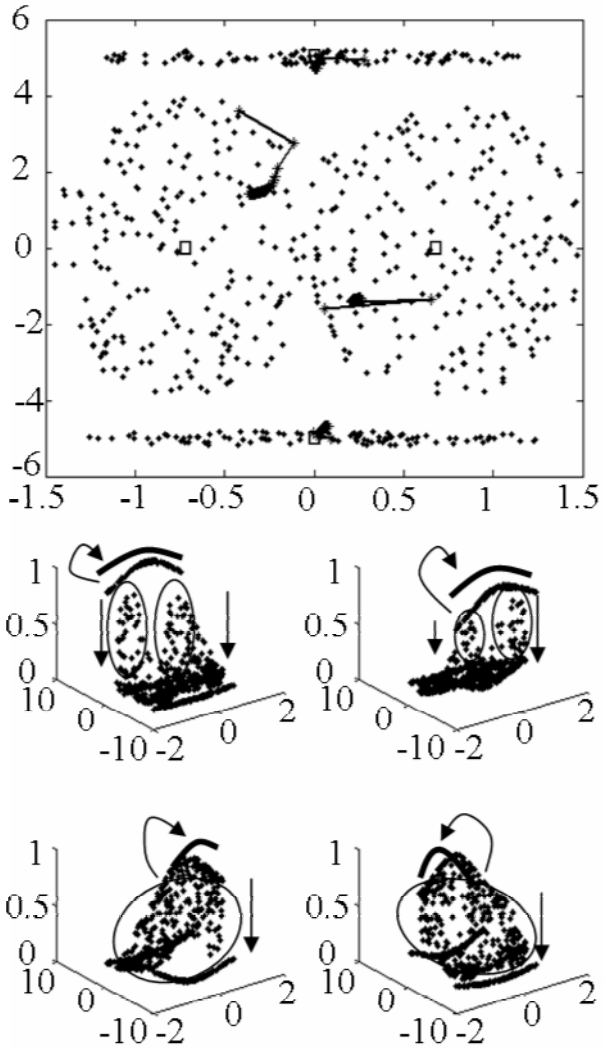

Рис.3. Результат определения параметров кластеров для примера 2

Для оценки выполненной кластеризации были вычислены показатели качества кластеризации (табл. 3).

Таблица 3. Показатели качества кластеризации, найденные алгоритмом

Наименование показателя	Обозначение	Номер примера	
		1	2
Коэффициент четкости	PC	0.7361	0,6648
Индекс четкости	CI	0.7356	0,6642
Энтропия	PE	0.5137	0,6338
Нормализованная энтропия	PE_n	0.5171	0,6380
Модифицированная энтропия	PE_m	0.3705	0,4572
Показатель компактности и изолированности	CS	0.7555	0,1155
Индекс эффективности	PI	3549	4371

4. Результаты

Для примера 1 роль эксперта заключалась в достаточно простых корректировках начальных условий для работы алгоритма самоорганизации.

Но если два кластера сближаются, то есть имеют общие точки, алгоритм самоорганизации не может найти истинные положения центров кластеризации (рис. 3). Использование предпочтений эксперта позволило для случая примера 2 достаточно быстро найти вид функций принадлежности для кластеров III и IV и разделить общие точки.

На рис. 3 предпочтения эксперта указаны стрелками, а внутри эллипсов заключены точки, которые, по мнению аналитика, нужно отнести к другому кластеру. В результате работы алгоритма значения функций принадлежности для кластеров I и II изменились незначительно. При этом значения функций принадлежности для кластеров III и IV изменились достаточно сильно (изменения указаны «кривыми» стрелками). С изменением значений функций принадлежностей были получены новые координаты центров кластеров III и IV. На рис. 3 очень хорошо

видно, что для кластеров III и IV «вершины» функций принадлежности сместились в сторону новых найденных центров. Следует также отметить, что вычислительные затраты при использовании технологии Visual Mining снизились на порядок, при этом точность полученных результатов была улучшена.

Литература

1. Батыршин, И.З. *Инвариантные кластерные процедуры, основанные на нечетком отношении сходства* / И.З. Батыршин, А.С. Климова // Сб. науч. трудов III-го Международного научно-практического семинара «Интегрированные модели и мягкие вычисления в искусственном интеллекте». М.: Физматлит, 2005. С. 119 – 125.
2. Bezdek (ed.) *Analysis of fuzzy information* / Vols. 1, 2, 3. CRC Press. 1987, P. 385.
3. Джоффрион, А. *Решение задач оптимизации при многих критериях на основе человеко-машинных процедур* / А. Джоффрион, Дж. Дайер, А. Файнберг // Вопросы анализа и процедуры принятия решений. М.: Мир, 1976. С. 126 – 145.