ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Б.Г. Миркин

МЕТОДЫ КЛАСТЕР-АНАЛИЗА ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ: ОБЗОР

Препринт WP7/2011/03
Серия WP7
Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2011

УДК 519.237.8
ББК 22.172
М63

М63 **Миркин, Б. Г.** Методы кластер-анализа для поддержки принятия решений: обзор : препринт WP7/2011/03 [Текст] / Б. Г. Миркин ; Национальный исследовательский университет «Высшая школа экономики». – М. : Изд. дом Национального исследовательского университета «Высшая школа экономики», 2011. – 88 с. – 150 экз.

В работе приведен обзор методов кластер-анализа, накопленных в литературе за 50 лет их активной разработки. При этом вместо обзора публикаций, которых имеются многие тысячи с самыми разнообразными алгоритмами и приложениями, в работе систематизированы задачи – по видам данных, видам кластерных структур, видам критериев и методов. Число подробно рассматриваемых методов относительно невелико. Зато большинство из них нетрудно перенести на другие виды данных и кластерных структур. Специальное внимание уделено принятию решений о параметрах алгоритмов, валидации и интерпретации. Во Введении систематизированы основные типы приложений для принятия решений. В Заключении обсуждаются основные нерешенные проблемы кластер-анализа. Бо́льшая часть текста написана так, чтобы человек с высшим техническим образованием мог быстро уяснить суть дела без углубления в технические детали.

УДК 519.237.8
ББК 22.172

Миркин Б.Г. – Лаборатория анализа и выбора решений; Отделение прикладной математики и информатики НИУ ВШЭ, Москва РФ; Department of Computer Science, Birkbeck University of London.

**Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp**

Содержание

Введение 4
1. Задачи и методы кластер-анализа 7
1.1. Типы данных 7
1.2. Вид искомой кластерной структуры 20
1.3. Критерии кластер-анализа 25
1.4. Методы кластер-анализа 33
2. Принятие решений в кластер-анализе 59
2.1. Число кластеров 60
2.2. Консенсус 63
2.3. Валидация кластеров 67
2.4. Наличие нескольких критериев 68
2.5. Интерпретация кластеров 69
3. Заключение: Основные направления развития 73
Литература 75

Введение

Кластер-анализ – растущая дисциплина (5,8 млн страниц из 404 млн, посвященных анализу по англоязычным запросам 4/2/11 в системе Google). Она с необходимостью должна быть связана с разработкой систем поддержки принятия решений (17 млн страниц из 199 млн, посвященных решениям по англоязычным запросам 4/2/11 в системе Google).

Под кластером обычно понимается часть данных (в типичном случае – подмножество объектов или подмножество переменных, или подмножество объектов, характеризуемых подмножеством переменных), которая выделяется из остальной части наличием некоторой однородности ее элементов. В простейшем случае речь идет о похожести элементов, в идеальном случае – о совпадающих значениях основных переменных или иного рода близости, выражаемой геометрической близостью соответствующих объектов.

Понятие принятия решений значительно более расплывчато. Международная литература относит сюда все системы, данные и знания, которые могут быть использованы при принятии технических, хозяйственных, экономических, политических или иных решений (см., например, англоязычный исторический обзор [Power 2007] или русскоязычный корпоративный обзор [Портал о корпоративных порталах 2010, http://corportal.ru]. В этом плане представляется, что кластер-анализ может играть ту же роль в принятии решений, что и классификация и систематизация в познании, выделяясь тем, что делается на основе более многообразной, обозримой и формализованной системы правил. Сюда прежде всего следует включать такие задачи поддержки принятия решений, как:

А. Структуризация данных

А1. Анализ состава и основных компонент данных

Например, при анализе группы поселений данного региона для целей социально-экономического мониторинга полезно знать основные типы поселений, количество типов, типообразующие признаки.

А2. Выявление групп объектов, к которым применимы одинаковые критерии

Это важно в таких задачах, как сопоставление регионов по уровню экономического развития, где, например, сельскохозяйственные

и промышленные регионы вряд ли целесообразно оценивать теми же самыми критериями.

А3. Выявление и анализ структуры взаимодействия основных подсистем

Интересно, например, исходя из структуры запросов между парами различных веб-сайтов попытаться выявить основные подструктуры и потоки запросов для дальнейшего использования в задачах планирования.

Б. Визуализация структуры системы

Визуализация остается одним из основных средств принятия поддержки решений, стимулирующим интуицию. После того как основные кластеры установлены, их взаимодействие и развитие значительно легче визуализировать, чем в исходной массе данных.

В. Выявление основных тенденций эволюции системы

Этот тип задач относительно новый, поскольку касается в основном информации о временных рядах и потоках данных. Более понятен одномоментный анализ объектов, находящихся на разных стадиях развития.

Г. Выявление связей между такими аспектами развития системы, как вход/выход, структура/функция, мотивация/действие и пр.

Интересно, например, проанализировать тенденции связей между входными переменными группы предприятий и их результатами. Для этого можно отдельно найти кластеры по входным переменным и кластеры по выходным переменным, а затем сравнить полученные структуры.

Несколько слов об истории кластер-анализа. Он зародился где-то перед Второй мировой войной как обобщение исследований:

(1) по выявлению и анализу популяций в экологии (индекс Чекановского, Вроцлавская таксономия);

(2) по построению внутренних факторов в психологии (в качестве «фактор-анализа для бедных», т.е. для тех, кто заменял трудоемкие операции обращения корреляционных матриц анализом структуры их элементов);

(3) «нумерической таксономии» в биологии, когда стало ясно, что не получается описать отдельными признаками все разнообразие видов.

Взрыв начался в 1960-е годы с внедрением вычислительной техники в университеты и оформлением направления сначала как под-

чиненного, «распознавания образов без учителя», а в последнее время независимого. В это время в СССР возникает несколько школ мирового значения, разрабатываются прикладные и теоретические работы, пишутся книги, опережающие время (см., например, [Авен и др. 1988; Айвазян и др. 1989; Айзерман и др. 1970; Браверман, Мучник 1983; Загоруйко 1999; Заславская, Мучник 1980; Мандель 1988; Миркин 1985] и др.; а также обзор [Mirkin, Muchnik 1996]. К сожалению, застой науки в постсоветскую эру эту работу в значительной степени разрушил. Направление предстоит возрождать на новом материале, новых приложениях и новых задачах.

В данной статье сначала будет приведен обзор методов кластер-анализа, накопленный в литературе за 50 лет их активной разработки. Вместо обзора публикаций, которых накопились тысячи, с самыми разнообразными алгоритмами и приложениями, мы систематизировали задачи по видам данных, видам кластерных структур и видам критериев. Самих методов мы рассматриваем не очень много, но зато большинство из них нетрудно перенести на другие виды данных и кластерных структур. Часть методов очень популярна, часть – не очень, но все-таки все основные популярные методы включены. Хотя мы старались, по возможности, описать все основные тенденции с учетом временных ограничений и субъективных предпочтений, конечно же, отдельные аспекты охарактеризованы более полно, чем другие. Другое замечание касается требований, предъявляемых к читателю: это должен быть человек с высшим техническим образованием, не обязательно имеющий опыт или склонность к анализу данных, но знающий основы анализа, линейной алгебры и теории множеств. Однако б*о*льшая часть обзора написана так, чтобы любой грамотный человек мог быстро уяснить суть дела без углубления в технические детали. В обзоре использованы материалы из недавних работ автора, прежде всего [Mirkin 2011] и [Mirkin, Muchnik 2008]. Другие взгляды на развитие кластер-анализа можно найти в недавних публикациях [Everitt et al. 2011; Jain 2010; Xu, Wunsch 2009; Garcia-Escudero et al. 2010].

Работа над обзором финансировалась в рамках проекта НИУ ВШЭ «Создание высокотехнологичного производства инновационных программно-аппаратных комплексов для эффективного управления

предприятиями и отраслями экономики современной России». Мы благодарны Лаборатории анализа и выбора решений НИУ ВШЭ за частичную финансовую поддержку работы, а ее руководителю Ф.Т. Алескерову – за побуждение к ее выполнению. Анонимный рецензент сделал замечания, учтенные при исправлении и доводке рукописи, особенно в части, касающейся стиля изложения и теоретических аспектов измерения.

1. Задачи и методы кластер-анализа

Обычно методы кластер-анализа разбивают на два типа: те, что ищут разбиения, и те, что ищут иерархии вложенных кластеров, однако это слишком грубая классификация, которая не может дать полного представления о богатстве методов, накопленных к настоящему времени.

Развивая предложенные нами ранее идеи [Mirkin 1996], мы предлагаем систематизировать методы кластер-анализа в разрезе следующих четырех компонент, однозначно определенных для любого метода кластер-анализа:

– тип данных;
– вид искомой кластерной структуры;
– критерий оценки кластерой структуры;
– метод построения кластерной структуры,

с тем, чтобы обратить внимание на методы, наиболее перспективные с той или иной точки зрения.

1.1. Типы данных

Когда говорят о данных, то имеет смысл отличать метаданные, т.е. названия объектов и признаков, шкалы измерения и пр., от самих данных, т.е. значений признаков на объектах и отношениях между ними. Имеет смысл различать следующие типы данных:

– таблицы объект – признак;
– матрицы сходства или близости;
– последовательности;
– неструктурированный текст;

– картографические данные;
– временные ряды.

Рассмотрим их в предложенной последовательности.

1.1.1. Таблицы объект – признак

Это наиболее базовый вид данных, ставший привычным в связи с распространением так называемых спредшитов типа «Эксел», поддерживающих данные в подобном виде, см. пример в нижеследующей таблице, взятой из учебника [Mirkin 2011].

Расшифровка признаков:

1) доход в миллионах долларов;

2) капитализация – стоимость всех активов в млн долл.;

3) Пос – число основных поставщиков;

4) ЭТ – «да», если используется электронная торговля; «нет», если не используется;

5) сектор – торговля, энергетика или промышленное производство.

Таблица 1. Компании: иллюстративные данные о восьми компаниях, характеризуемых пятью признаками. Таблица дополнительно разделена на три горизонтальные секции, отражающие основную продукцию компаний, заложенную также в их названиях: первые три компании производят в основном продукты группы A, следующие три – продукты группы B, две последние – продукты группы C

Название компании	**Доход, млн долл.**	**Капитализация, млн долл.**	**Пос**	**ЭТ**	**Сектор**
Aversi	19.0	43.7	2	Нет	Энергетика
Antyos	29.4	36.0	3	Нет	Энергетика
Astonite	23.9	38.0	3	Нет	Промышленность
Bayermart	18.4	27.9	2	Да	Энергетика
Breaktops	25.7	22.3	3	Да	Промышленность
Bumchist	12.1	16.9	2	Да	Промышленность
Civok	23.9	30.2	4	Да	Торговля
Cyberdam	27.2	58.0	5	Да	Торговля

Несмотря на иллюстративный характер, данные табл. 1 позволяют использовать кластер-анализ для ответа на такие осмысленные вопросы поддержки принятия решений, как:

– Как визуализировать компании точками плоскости так, чтобы близость точек отражала сходство признаков?

– Будут ли кластеры компаний по признакам отражать группы основных продуктов (которые не входят в число признаков)?

– Если да, можно ли предложить формальное правило для соотнесения продуктов компаниям, которое можно было бы применять и к другим, не вошедшим в таблицу компаниям?

– Признаки табл. 1 распадаются на две группы, структурные (ЕТ, Сектор, Пос) и рыночные (доход и капитализация): Существует ли какое-либо отношение между этими группами?

1.1.1.А Шкалы измерения; бинарные признаки как количественные

В табл. 1 встречаются три основных типа шкал: количественная, бинарная и номинальная. Для целей обработки удобно приводить бинарные и номинальные шкалы к количественному виду, кодируя «да» единицей, а «нет» нулем. При этом каждое значение номинального признака «Сектор» рассматривается как отдельная бинарная переменная, дающая ответ на соответствующий вопрос: Сектор энергетики? Промышленности? Торговли? При этом перекодированная таблица будет иметь вид табл. 2.

Следует отметить, что не все принимают такую количественную перекодировку. Многие авторы считают, что для разношкальных данных необходима не перекодировка, а перевод в формат матриц близости или сходства между объектами (см. ниже в разд. 1.2). На наш взгляд, это связано с отсутствием понимания того, что количественная перекодировка – это не эвристическое, т.е. ничем кроме личного вкуса не обоснованное, а теоретически состоятельное преобразование.

Таблица 2. Данные табл. 1, переведенные в количественный формат

Номер	Доход	Кап.	Пос	ЭТ	Энер	Пром	Торг
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1

Она соответствует принципам так называемой репрезентационной теории измерений [Суппес, Зинес 1967; Пфанцагль 1990]. Согласно этой теории, идентичность количественных признаков не меняется при всевозможных интервальных (или аффинных) преобразованиях, связанных с изменением точки отсчета и масштаба. В неопубликованной работе автор доказал, что это то же самое, что допускать осмысленное вычисление и сравнение средних значений. Для частного случая, когда усредняющая функция множества чисел $X = \{x_1, x_2, ..., x_n\}$ имеет вид обобщенного среднего по Колмогорову, $f(x_1, x_2, ..., x_n) = H^{-1}(\Sigma a_i H(x_i))$, этот результат был установлен Орловым (1985). Очевидно, что бинарные признаки, будучи номинальными, допускают любые взаимно-однозначные преобразования. Однако каждое такое преобразование определяется ровно двумя параметрами в силу бинарности признака, которые, таким образом, связываются с двумя параметрами интервального преобразования. Средние значения бинарных признаков – частоты встречаемости соответствующих категорий – основное средство анализа нечисловых данных. Таким образом, эти два вида шкал могут вполне мирно сосуществовать.

К сожалению, этого нельзя сказать о так называемых порядковых, или ранговых, шкалах, часто используемых в методологии принятия решений. Такие шкалы определены с точностью до произвольного монотонного, а не интервального преобразования, что связано с более сложными математическими образованиями, конусами, а не линейными пространствами, как в случае количественных и бинарных данных. В 70–80-е годы XX в. французские специалисты пытались объединить эти два типа математических структур в единую теорию, но неудачно. Так и стоят ранговые шкалы особняком.

1.1.1.Б Данные с независимыми столбцами и их стандартизация

Даже и в количественном формате возможны разные виды данных с точки зрения их моделирования и предварительной стандартизации. Формат вида табл. 2 – так называемый с независимыми столбцами – разные признаки несоизмеримы. Для приведения признаков к соизмеримому виду каждый стандартизуется по-своему, путем изменения точки отсчета и перемасштабирования по формуле

$$y_{iv} = (x_{iv} - a_v)/b_v \qquad (1.1)$$

Здесь $X = (x_{iv})$ обозначает исходную матрицу данных, а $Y = (y_{iv})$ – стандартизованную. При этом $i \in I$ – объекты, а $v \in V$ – признаки. Параметр a_v задает сдвиг точки отсчета, а b_v – новый масштаб для каждого признака $v \in V$.

После преобразования нулевая точка 0 = (0,0, ..., 0) приобретает уникальное значение, поскольку любое линейное преобразование AY геометрически выражает косоугольное вращение осей с изменением их масштабов, оставляющее точку отсчета 0 инвариантной. Метафорически точка отсчета может быть уподоблена «глазу», из которого обозреваются данные. Поэтому для целей анализа данных, включая кластер-анализ, начало координат лучше всего помещать где-то в районе центра множества точек, представляющих объекты.

Что касается масштабирующих коэффициентов b_v, их следует выбирать исходя из идеи выравнивания относительных весов признаков. Несмотря на то что интуитивно принцип довольно понятен, он не нашел в общем контексте анализа данных разумного воплощения. Кажущееся очевидным использование для этой цели стандартного отклонения, переводящее все признаки к «единому» масштабу единичного стандартного отклонения, представляется скорее вредным с точки зрения выявления кластер-структуры. Это иллюстрируется примером двух признаков, имеющих одну и ту же область определения, но отличающихся плотностями распределения (рис. 1).

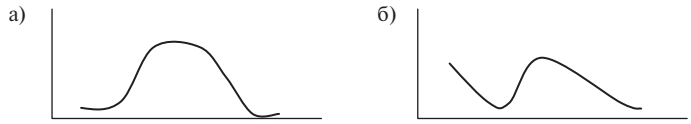

Рис. 1. Сравнение одномодальной плотности (а) с двумодальной плотностью (б) – во втором случае стандартное отклонение больше, что делает второй признак менее значимым, чем первый, при нормализации стандартными отклонениями

Поскольку для двумодального распределения стандартное отклонение значительно больше, все его значения сильно уменьшатся по сравнению со значениями одномодального признака. Тем самым при вычислении расстояний этот признак будет доминировать, показывая одномодальное распределение, соответствующее ситуации однородности объектов. При этом значение второго признака, по кото-

рому объекты распадаются на два кластера, сильно уменьшится. С точки зрения кластер-анализа, веса признаков должны быть обратными. Объяснение представленного явления лежит на поверхности (см. [Mirkin 2005]). Дело в том, что в понятии стандартного отклонения смешаны две разнородных характеристики – масштаб измерения признака и форма распределения. С точки зрения кластер-анализа, для нормализации лучше использовать – и это доказано многократными экспериментами (см. [Milligan 1980; Steinley, Brusco 2007]) – размах или полуразмах признаков. Размах – это разность между максимумом и минимумом значений для небольшого объема данных или между 99% и 1% квантилями в случае больших объемов (для уменьшения эффекта случайных выбросов). Любопытно, что нормализация на размах является адекватной с точки зрения абстрактной теории измерений, а на стандартное отклонение – нет. Действительно, отношение (1.1) не изменится, если признак x подвергнуть аффинному преобразованию $x' = \alpha x + \beta$, в случае, когда в знаменателе находится размах. В случае же когда знаменатель – стандартное отклонение или любая другая статистика, не вычитающая значений, то величина (1.1), вообще говоря, изменится при аффинном преобразовании признака.

Относительно выбора сдвига точки отсчета существует несколько популярных точек зрения. В распознавании образов обычно берут середину интервала размаха признака. Другие, включая автора, используют среднее арифметическое значение, тем более что этому имеются определенные теоретические подтверждения, связанные с совместной обработкой количественных и номинальных признаков (Mirkin 2005, 2011).

1.1.1.B. Сравнимые и суммируемые данные, их стандартизация, бинарные данные

В некоторых случаях значения в разных столбцах имеет смысл сравнивать, а иногда даже и суммировать. Пример первого случая – данные психологических экспериментов, в которых респонденты (строки) оценивают уровень своего отношения к различным ситуациям (столбцы) в шкале, скажем, от 0 (не хочу участвовать) до 10 (самая желательная). Подобный характер имеют и данные по экспрессии генов, где по строкам находятся различные биомолекулярные образования, по столбцам – ткани организма, а значения – уровни

экспрессии, измеряемые уровнем проявления того или иного красящего вещества. Эти уровни можно сравнивать между столбцами, не только внутри них, а значит и стандартизация должна носить тотальный характер – сдвиг всех элементов таблицы на одну и ту же величину, скажем, среднее по всей таблице. Нормализация на одну и ту же величину, обычно не имеет смысла, так как не влияет на результаты анализа структуры.

В некоторых случаях имеет смысл не только сравнивать табличные значения, но и суммировать их, например, когда речь идет о семьях или группах населения (строки), а по столбцам идут денежные расходы на различные нужды, или же речь может идти о парном распределении населения региона по профессиям (столбцы) и поселениям (строки) – такие данные называются таблицами сопряженности в статистике. Эти величины удобно суммируются при объединении поселений и/или профессиональных групп, и могут поэтому интерпретироваться как части одного и того же «потока». В этом случае удобно преобразовывать данные в относительный формат как по «источнику» (строка), так и «стоку» (столбец). В принципе, эти данные можно считать измеренными в шкале отношений, как любые потоковые данные, т.е. с точностью до единицы измерения (население – в миллионах или тысячах, денежную массу – в рублях или евро, массу – в тоннах или центнерах).

В частности, если обозначить численность населения в клетке $(k, l) \in K \times L$ через N_{kl}, то доли определятся отнесением этих величин к суммарной численности населения N (имеется в виду, что категории, формирующие строки (или столбцы), не пересекаются и покрывают все население). Тогда доли $p_{kl} = N_{kl} / N$ могут суммироваться как внутри строк, так и внутри столбцов, формируя величины $p_{k+} = \Sigma_l p_{kl}$ и $p_{+l} = \Sigma_k p_{kl}$ так называемых маргинальных распределений. Относительная разница между долей l в строке k, $P(l/k) = p_{kl}/p_{k+}$ и долей $P(l)$ столбца l во всем населении и образует безразмерный коэффициент Кетле

$$q(l/k) = [p_{kp}/p_{k+} - p_l]/ p_{+l} = \frac{p_{kl}}{p_{k+} p_{+l}} - 1, \qquad (1.2)$$

который оказывается удобным способом стандартизации суммируемых данных (Benzecri 1989, Mirkin 2011).

Важным случаем сравнимых/суммируемых данных являются таблицы бинарных данных, в которых все элементы либо нули, либо

единицы. Такие данные, с одной стороны, характеризуют чисто качественную информацию о наличии или отсутствии связи между строкой и столбцом, а, с другой стороны, при их количественной интерпретации, объясненной выше, могут рассматриваться как сравнимые или даже суммируемые данные, с применением соответствующих формул для их стандартизации.

1.1.2. Матрицы сходства или близости

Такие данные имеют вид матриц связи $A = (a_{ij})$, где $i, j \in I$, или расстояний (близости) $D = (d_{ij})$, $i, j \in I$, где a_{ij} выражает степень связи (d_{ij} – величину расстояния) между объектами и $i, j \in I$. То есть строки и столбцы здесь не независимы, а соответствуют одному и тому же объекту $i \in I$. При этом объекты i и j тем более похожи друг на друга, чем больше величина связи и чем меньше величина расстояния между ними. Понятно, что все элементы такой матрицы сравнимы по всей таблице, а в случае, когда они характеризуют потоки (людей, денег, товаров), то и суммируемы. Информация, содержащаяся в графах, взвешенных графах и сетях, очевидно представляется подобным же образом. При этом все сказанное в разделе 1.1 о стандартизации таких данных относится и к матрицам сходства или близости. Особенно привлекательной становится опция сдвига шкалы измерения степени связи как параметр, имеющий разумный смысл характеристики порога значимости связи при принятии решений и управляемый пользователем.

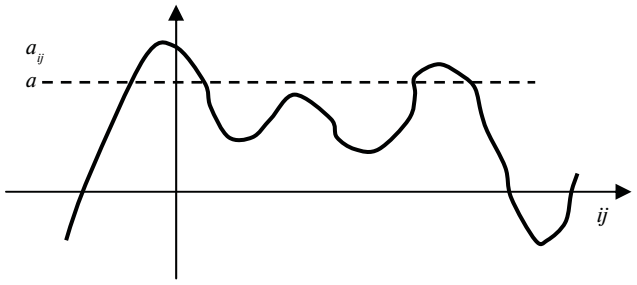

Рис. 2. Иллюстрация эффекта вычитания порога a из значений матрицы связи. На оси абсцисс условно отражены пары объектов (i, j) тогда как ось ординат соответствует степени связи. Штриховая линия показывает уровень связи, остающиеся положительными после вычитания порога a два островка представляют большой контраст по сравнению с исходным монолитом

В самом деле, имея в виду, что кластеры должны состоять из близких объектов, пользователь может установить порог индивидуальной связи a при связи a_{ij} ниже которого объединение объектов i и j в один кластер нежелательно. Вычитание величины a из всех элементов матрицы связи произведет эффект, проиллюстрированный на рис. 2 – ось абсцисс как бы поднимется до уровня a, так что объединяться в кластеры будут только «разрозненные» островки, где функция связи остается положительной.

Матрицы связи раньше генерировались в основном в психологических экспериментах, а матрицы близости – как матрицы расстояний между объектами по данным типа объект – признак. В настоящее время появились три новых источника таких данных:

1) Матрицы связей возникают при наложении друг на друга объектов сложной природы – изображений, протеиновых последовательностей, неструктурированных текстов и т.п.

2) Матрицы связи кодируют данные о социальных и иных сетях в Интернете, как, например, число или продолжительность обращений с одной страницы на другую.

3) При анализе изображений (сегментация и т.п.) оказалось необычайно эффективным преобразование Евклидовых расстояний в матрицы сходства с помощью так называемых кернел-функций (были введены в ИПУ РАН под названием потенциальных функций, см. [Браверман, Мучник 1983]), порождающих положительно полуопределенные матрицы. Основное используемое преобразование – Гауссиан

$$a_{ij} = \exp\{-d^2(x,y)/2\sigma^2\}, \qquad (1.3)$$

где $d^2(x, y)$ – квадрат Евклидова расстояния между векторами x и y. В настоящее время свойства этого преобразования активно исследуются, поскольку они не только эффективны в практической работе, но и в теоретическом плане, так как позволяют интегрировать в едином подходе многие конструкции, до недавнего времени рассматривавшиеся изолированно (оценка функций плотности данных, преобразование переменных, критерии кластер-анализа и пр., см., например, [Jenssen et al. 2005, 2008]).

1.1.3. Последовательности

Этот вид данных стал популярен сравнительно недавно, прежде всего в связи с появлением данных о протеинах как последователь-

ностей в алфавите 20 аминокислот, а также коротких текстов на естественном языке – заголовков, СМС-сообщений, и пр. Последовательность – объект сложной структуры, позволяющей формировать признаки, комбинируя структурные элементы, что в настоящее время недостаточно исследовано. Можно указать четыре наиболее популярных способа их предварительной обработки.

А. Попарное наложение (выравнивание)

Строки располагаются друг под другом так, чтобы максимизировать число позиций, в которых расположены совпадающие символы. При этом разрешается возможность создания разрывов в любой из последовательностей, что существенно увеличивает возможности выравнивания и сложность алгоритмов. Исходя из гипотезы, согласно которой последовательности совпадали при их возникновении, а расхождение произошло в процессе исторического развития или редактирования, разрыв в одной из последовательностей может интерпретироваться либо как выпадение символов (делеция) в ней самой, либо как включение (инсерция) символов в сравниваемую последовательность. Подобный подход характерен прежде всего для биоинформатики. Соответственно, было разработано несколько мер связи между последовательностями как вероятностей их происхождения из единого предка, тогда как в анализе текстов используют так называемое редакционное расстояние В. Левенштейна (1965). Результирующая мера близости или сходства порождает соответствующую матрицу расстояний или связи (см. п. 1.2), которая в последующем и обрабатывается.

Б. Формирование профиля и скрытой марковской модели

Для этого производится множественное наложение последовательностей (или выравнивание), опять же с возможностью разрыва. Затем берется максимально возможное множество позиций, в которых участвуют все рассматриваемые последовательности, а остальные позиции, с данными только от части последовательностей, выбрасываются. Этот фрагмент затем агрегированно представляется в виде профиля – совокупности позиций, каждой из которых приписано распределение символов: если, скажем, имеется всего пять символов, причем в данной позиции A участвует в 20% последовательностей, B – в 50%, и C – в 30%, то распределение имеет вид вектора (0.2, 0.5, 0.3, 0.0, 0.0). Более информативное представление – так на-

зываемая скрытая марковская модель (СММ). Согласно этой модели, последовательности генерируются переходом конечного автомата из состояния в состояние, причем при каждом переходе с фиксированными для каждого состояния вероятностями испускаются символы, формирующие последовательность [Karplus et al. 1998]. Это позволяет оценивать вероятность любой последовательности и в конечном счете степень ее принадлежности данному множеству.

В. Признаковое описание

Пользователь фиксирует определенное множество признаков (или предикатов), в простейшем случае – «ключевых слов», которые рассматриваются как столбцы формируемой таблицы данных, в данном случае – бинарной, где 1 ставится тогда, когда ключевое слово входит в последовательность, и 0 – когда нет. Это может быть также несколько более информативный признак – частота вхождений данного ключевого слова. Частоты допускают дальнейшее перекодирование. В частности, очень популярно так называемое tf-idf (term frequency – inverse document frequency) кодирование, при котором частота вхождения ключевого слова в последовательность делится на логарифм числа последовательностей, в которые это слово входит. Таким способом отфильтровывают слишком часто встречающиеся ключевые слова, поскольку их информационная ценность невелика.

Для дальнейшей обработки, как подчеркивают некоторые авторы, большое значение имеет дальнейшая нормировка (Евклидова) строк, соответствующих объектам [Dhillon, Modha 2001; Ng et al. 2002] – после нормирования скалярное произведение выражает косинус угла между объект-строками.

Г. Представление фрагментами

Каждая последовательность представляется подмножеством некоторых ее фрагментов. Исходно использовались так называемые n-граммы – совокупность всех неразрывных фрагментов длины n. Очевидно, в последовательности $x_1, x_2, \ldots x_N$ длины N, общее число n-грамм равно $N - n + 1$ – по числу возможных начальных символов $x_1, x_2, \ldots, x_{N-n+1}$. Позднее стали использоваться совокупности n-грамм с n, заключенным в заданном интервале, скажем, от 3 до 7. Наконец, удачная и обозримая структура, становящаяся все более популяр-

ной – суффиксное дерево [Gusfield 1997] – представляет все фрагменты последовательности. В работе [Pampapathi, Mirkin, Levene 2006] использованы суффиксные деревья, аннотированные частотами встречаемости соответствующих фрагментов (АСД) (рис. 3).

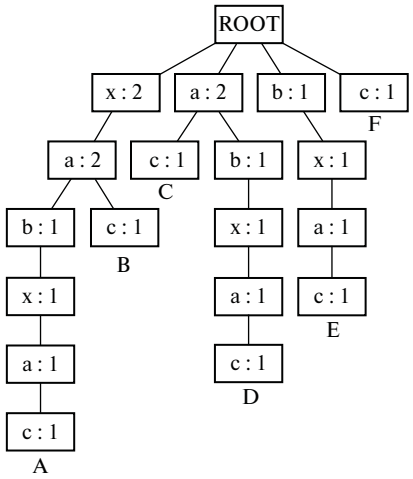

Рис. 3. Аннотированное суффиксное дерево для последовательности *xabxac*; ее суффиксы (конечные фрагменты) соответствуют путям от корня к вершинам, причем каждая вершина аннотирована частотой встречаемости соответствующего фрагмента. Мы видим, что нетривиальная частота 2 приписана только одной нетривиальной последовательности, *xa* (заимствовано из [Черняк 2010])

Наложение двух АСД осуществляется довольно просто и позволяет оценить степень сходства соответствующих последовательностей, а также определить фрагменты, дающие наибольший вклад в сходство [Pampapathi 2008; Черняк 2010].

1.1.4. Неструктурированный текст

Неструктурированный текст можно рассматривать как совокупность последовательностей, составляющих законченные предложения, поэтому здесь применимы приемы, описанные в предыдущем разделе. Особенно перспективны (a) Признаковое описание, включая кодирование в системе tf-idf, и (b) Представление фрагментами, особенно аннотированные деревья. Разработки, связанные с так называемой Обработкой естественных языков (NPL), включая грам-

матический разбор предложений, должны бы сыграть свою роль, что пока остается делом будущего, как и представление предложений и текстов в семантически насыщенной системе типа «что-где-когда-почему» [Feldman, Sanger 2007; Manning 1999].

1.1.5. Картографические данные

Картографические данные – изображения, географические информационные системы, океанические исследования и другие приложения ассоциированы с информацией, расположенной в узлах двумерной целочисленной решетки (например, пикселах экранного поля), в которой с необходимостью возникают так называемые пространственные корреляции – данные в близких узлах имеют близкие значения. И вообще, топология двумерной решетки предполагает, что картографические кластеры должны состоять из близко расположенных точек. До недавнего времени подобные соображения учитывались исключительно с помощью навязываемых извне ограничений типа того, что «если два узла включены в кластер, то находящиеся между ними узлы тоже должны быть включены». В последнее время появилась надежда, что пространственные связи могут быть автоматически учтены в тех или иных методах. С этой точки зрения определенные надежды связаны с так называемым Лапласовым преобразованием (см. разд. 1.4.2.3). Еще более адекватным представляется применение здесь так называемого многоагентного подхода, при котором узлы решетки наделяются определенной свободой примыкать к тому или иному кластерному агенту, а агенты, в свою очередь, действуют на решетке в рамках определенных пространственных окон и могут рекрутировать узлы по тому или иному критерию. В целом картографические данные пока недостаточно осознаны как специальный объект анализа (см., например, [Bivand et al. 2008]), несмотря на определенные успехи в развитии Географических информационных систем (ГИС) (см. [Lo, Yeung 2006]).

1.1.6. Временные ряды

Данные любого формата можно рассматривать в динамике. Развитие спутниковых средств связи и Интернета сделали возможным сбор и поддержку огромных массивов временных данных, прежде всего картографического типа, а также рядов различного рода действий, связанных с Интернетом – так называемый поток данных.

Наука фактически оказалась не готовой к такому изобилию, если не считать относительно частной задачи сегментации отдельного временного ряда на относительно стационарные фазы. Поэтому в настоящее время проблематика анализа временных рядов, особенно многомерных, находится в самом начале своего развития [Mitsa 2010]. Еще менее развиты представления, относящиеся к анализу потока данных [Gama 2010].

1.2. Вид искомой кластерной структуры

Обычно выделяют два вида структур: разбиения и бинарные иерархии. Однако на самом деле различимых кластерных структур, пользующихся определенной популярностью, значительно больше:

1. Множества центроидов.
2. Разбиения.
3. Разбиения с центроидами.
4. Иерархии.
5. Отдельные кластеры.
6. Аддитивные кластеры.
7. Разбиения со структурой.
8. Бикластеры.
9. Нечеткие кластеры и разбиения.
10. Смеси распределений.
11. Самоорганизующиеся карты Кохонена.

Далее эти виды структур будут кратко охарактеризованы, прежде всего с точки зрения оснований.

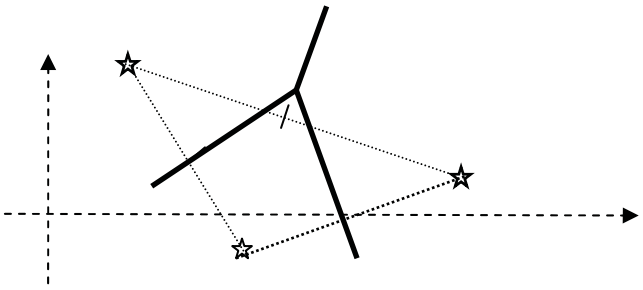

Рис. 4. Диаграмма Вороного для системы трех центроидов, представленных пятиконечными звездами

1. Множества центроидов

Множества центроидов возникли в исследованиях по нейронным системам, прежде всего сетей Кохонена, который называл их learning quantization vectors. Как только задано конечное множество центроидов в признаковом пространстве, каждая точка пространства приписывается одному из центроидов согласно так называемому принципу минимального расстояния – ближайшему в рассматриваемой метрике, обычно Евклидовой. При этом возникает так называемая диаграмма Вороного – русского математика XIX в. – совокупность гиперплоскостей, разделяющих области притяжения каждого из центроидов (рис. 4). Очевидно, эти области притяжения образуют разбиение пространства, определяемое данной системой центроидов.

2. Разбиения

Разбиение – совокупность непустых непересекающихся классов – одна из самых популярных кластерных структур, особенно часто применяемая при анализе данных о сходстве между объектами. Типичная проблема, возникающая при этом – интерпретация классов получаемого разбиения. Поэтому по возможности кластеры сопровождаются их «представителями» – объектами или усредненными характеристиками, представляющими основные тенденции кластера.

3. Разбиения с центроидами

Именно такова структура, в которой представлены и кластеры, и их центроиды. Дополненная принципом минимального расстояния, она всегда будет представлена выпуклыми, а значит и хорошо интерпретируемыми кластерами – метод κ-средних, использующий эту структуру, недаром является самым популярным методом кластер-анализа.

4. Иерархии

Кластерная иерархия на множестве объектов I – это совокупность H вложенных подмножеств S, называемых кластерами, и удовлетворяющая свойству: при любых $S1$ и $S2$ из H их пересечение $S1 \cap S2$ либо пусто, либо совпадает с одним из них. Такие иерархии получают либо путем агломерации – объединения более мелких кластеров, либо путем разделения более крупных кластеров на более мелкие. И те, и другие обычно получаются путем дихотомий, т.е. являются бинарными. Раньше такие структуры использовались почти исключительно для

порождения разбиений. В настоящее время они начинают использоваться и сами по себе, прежде всего как способ хранения данных. Например, основная структура данных при анализе изображений – так называемые quadtrees, четвертичные деревья, получается последовательным разделением прямоугольной решетки пикселов на четыре равные прямоугольника. Вероятно, по мере развития иерархических таксономий и онтологий как способов хранения и поддержки знаний иерархии будут все чаще использоваться сами по себе.

5. Отдельные кластеры

Эта структура оправдывает себя в тех частых ситуациях, когда часть объектов «не ложится» в кластеры, будучи либо уникальными, включая выбросы и ошибки, либо частями бесформенной массы, обозначаемой в перечислениях «и другие».

6. Аддитивные кластеры

Аддитивные кластеры – это совокупность отдельных кластеров, обычно пересекающихся, в которой каждый кластер ассоциирован с положительной величиной – интенсивностью кластера. Предполагается, что сходство между любыми двумя объектами равно сумме интенсивностей тех кластеров, которым принадлежат оба объекта. Понятие хорошо описывает семантические связи, которые можно считать суммами связей, возникающих в соответствии с «элементарными» смыслами в данной культуре (см. [Фрумкина, Миркин 1986]. Понятие возникло в психологии США [Shepard, Arabie 1979] и еще раньше в СССР – в работах по аддитивному разложению матриц связи [Миркин 1976].

7. Разбиения со структурой

Для данного множества объектов I эта конструкция определяется графом (S, τ), где S – множество K агрегированных вершин, каждая из которых соответствует подмножеству $S_k \subset I$ исходных объектов ($k = 1, ..., K$), а τ – множество дуг, представляющих скопления дуг исходного графа. Мучник (1974) и его коллеги рассматривали ситуацию, когда τ задано заранее, а множества S_k могут пересекаться. Миркин (1974), напротив, считал структуру τ не заданной, а искомой, но зато S_k обязаны были не пересекаться. Разбиение со структурой отражает структуру большого графа и использовалось прежде всего для анализа организационных структур. По идее, в настоящее время,

когда социальные интернет-сети становятся явлением повседневной жизни, данное понятие должно быть востребованным. Однако это не так. В 1970–1980-е годы в международной социометрике возникло соответствующее понятие – блок-модель, но по непонятной причине оно не получило компьютерной поддержки и в настоящее время не используется.

8. Бикластеры

Это понятие было введено в книге [Миркин 1996] (отражая работы, начатые еще в статье [Hartigan 1972] и в настоящее время широко используется в анализе данных. Понятие «бикластеры» предполагает наличие двух множеств объектов: одно, относящееся к строкам, а другое – к столбцам матрицы данных – например, множество предприятий и множество характеристик x производственной деятельности, или множество генных продуктов и множество клеточных тканей, в которых регистрируются уровни их экспресии. Бикластер – это пара подмножеств (V, W), соответствующая в определенном смысле однородным предприятиям и показателям, или генным продуктам и тканям, которые проявляют однородную экспрессию. Аналогично определяются бикластерные разбиения и иерархии.

9. Нечеткие кластеры и разбиения

Все вышеопределенные структуры могут быть перенесены в область нечетких множеств, что особенно полезно в ситуациях, когда речь идет об анализе непрерывных или ментальных конструкций, в которых сама природа вещей не позволяет провести определенных четких границ. Напомним, что нечеткое множество на множественосителе I определяется как неотрицательная функция μ, такая что $\mu(i) \geq 0$ для всех $i \in I$, причем $\mu(i) \subset 1$ для всех i, и $\mu(i) = 1$ интерпретируется как четкая принадлежность. Нечеткое разбиение – это совокупность таких нечетких множеств μ_k, что $\Sigma_k \mu_k(i) = 1$ для всех $i \in I$.

10. Смеси распределений

Данные финансовых потоков или астрономических наблюдений могут рассматриваться как случайная выборка из бесконечной популяции. В этом случае уместна модель смеси распределений, в которой каждый κ-й кластер представлен одномодальной функцией плотности $f(x, \alpha_k)$, обычно Гауссовой (рис. 5), хотя в последнее время все чаще можно встретить так называемые распределения Дирихле.

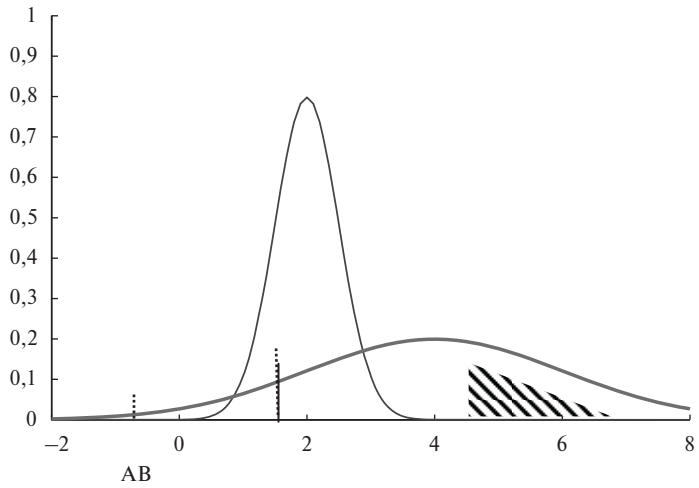

Рис. 5. Два Гауссовых кластера с сильно отличающейся кривизной порождают проблему в интерпретации принадлежности объектов.
Интервал (*A*, *B*) – единственное место, где остроконечный кластер более вероятен, чем второй, более пологий. Четкий кластер, соответствующий пологой функции плотности, получается разрывным, что противоречит интуитивному содержанию понятия кластера

Гауссова плотность имеет формулу, определяемую средним вектором m_k и матрицей ковариации Σ_k:

$$f(x, m_k, \Sigma_k) = \exp(-(x - m_k)^T \Sigma_k^{-1} (x - m_k) / 2) \Big/ \sqrt{(2\pi)^V |\Sigma_k|}. \quad (1.4)$$

Гауссовы кластеры эллипсоидальны, так как поверхность постоянства функции плотности определяется уравнением $(x - m_k)^T \Sigma_k^{-1} (x - m_k) = c$, где c – константа. Средний вектор m_k определяет местоположение k-го кластера.

Согласно модели смеси распределений, наблюдаемые векторы считаются независимой случайной выборкой из популяции с функцией плотности $f(x) = \sum_{k=1}^{K} p_k f(x, \alpha_k)$, где $p \geq 0$ – вероятности индивидуальных кластеров такие, что $\sum_{k=1}^{K} p_k = 1$.

11. Самоорганизующиеся карты Кохонена

Самоорганизующиеся карты Кохонена – это подход, ориентированный на непосредственную визуализацию объектов в целочисленной решетке, размер которой определяется пользователем (рис. 6, где изображена карта Кохонена размером 7×12). Этот подход использует топологию решетки, наряду с топологией признакового пространства и в конечном счете размещает близкие объекты в близких клетках решетки, как показано на рис. 6. Этот подход был очень популярен в 1990-е годы и только поэтому получил отражение в этом тексте. В настоящее время его популярность падает, и, вероятно, в скором времени карты Кохонена перестанут применять вообще. Основной недостаток метода – отсутствие средств интерпретации.

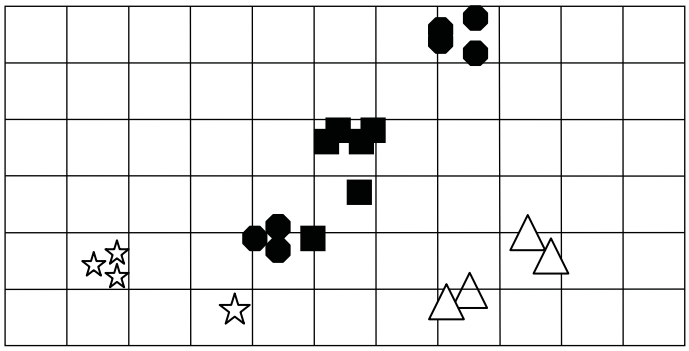

Рис. 6. Иллюстрация карты Кохонена и результата представления объектов с использованием геометрических фигур для указания сходства

1.3. Критерии кластер-анализа

Основная проблема кластерных критериев – отсутствие четкого обоснования, за исключением случаев, когда речь идет о чисто технических задачах типа модуляризации больших схем. Эти критерии пока не удается математически связать с главной целью кластер-анализа – улучшением понимания структуры данных, а через них – изучаемого явления или процесса, которая пока не может быть уточнена до математически ясной постановки.

Поэтому так важны исследования по валидации критериев – либо через анализ различных конкретных данных, либо через установление связей с другими методами и подходами.

В последнее время в кластер-анализе используются только оптимизационные критерии, хотя вначале развивались методы, основанные на математических определениях понятия «кластер», а также чисто эвристические методы, в которых каждый шаг построения кластеров соответствовал интуитивному представлению о них (обзоры такого рода работ можно найти в книгах Загоруйко (1999), Манделя (1988) и Миркина (1985)).

Оптимизационные критерии кластер-анализа могут быть разделены на три типа:

(а) эвристические; в таких критериях формализуется интуитивная идея, что объекты внутри кластеров должны быть близки друг к другу, а в разных кластерах – далеки друг от друга;

(б) аппроксимационные; такие критерии основаны на представлении искомой кластерной структуры математическими объектами того же типа, что и данные, обычно в виде матриц, так что в качестве критерия выступает степень близости между матрицей исходных данных и матрицей формируемой кластер-структуры.

(в) статистического оценивания; обычно это критерий максимального правдоподобия какой-либо статистической модели, такой, как смесь распределений.

В настоящее время основное значение имеют эвристические критерии, которые, по мере их использования в анализе данных, постоянно модифицируются и уточняются, в том числе на основе аппроксимационных или статистических соображений.

Рассмотрим примеры эвристических критериев для двух основных в данном обзоре видов данных: (А) матрицы связи и (Б) матрицы объект – признак.

1.3.А Критерии разбиения по матрице связи

При заданной матрице связи $A = (a_{ij})$ на множестве I будем, для простоты, делить I на две части, S_1 и S_2, так, чтобы связь между S_1 и S_2 была бы минимальной, а внутри этих множеств – максимальной. Естественная формализация такого критерия может быть выражена в терминах суммарной связи. Обозначим связь между S_f и S_g через $a(S_f, S_g) = \sum_{i \in S_f} \sum_{j \in S_g} a_{ij}$. Тогда $a(S_1, S_1)$ – суммарная связь внутри S_1, а $a(S_1, S_2) = a(S_2, S_1)$ – суммарная связь между S_1 и S_2 (A предполага-

ется симметричной). Поскольку сумма $a(S_1, S_1) + a(S_2, S_2) + a(S_1, S_2) + a(S_2, S_1)$ равна сумме $a(I)$ всех связей и, значит, постоянна, то естественный критерий минимизации $a(S_1, S_2)$ (минимальный разрез) одновременно максимизирует сумму связей внутри кластеров:

$$aw(S_1,S_2) = a(S_1, S_1) + a(S_2, S_2), \tag{1.5}$$

что является удачным свойством критерия. Тем не менее этот критерий не работает для наиболее часто встречаемых, неотрицательных матриц связи, так как он приводит к очевидному – и тривиальному – оптимальному решению: собрать все объекты в один кластер, выделив в другой кластер только один объект – тот, у которого самые слабые связи с остальными. Такое несбалансированное решение, конечно, не может рассматриваться по-настоящему кластерной структурой, что приводит к необходимости модификации критерия. Такая модификация была предложена Дорофеюком и Браверманом (см. [Браверман, Мучник 1983], которые предложили делить сумму внутренних связей кластера на его численность, так что максимизируемый критерий превращается в

$$ab(S_1,S_2) = a(S_1, S_1)/|S_1| + a(S_2, S_2)/|S_2|. \tag{1.6}$$

Последний критерий приводит к более сбалансированным кластерным структурам. Позднее Миркин и Мучник (1981) показали, что этот критерий возникает также и в аппроксимационных задачах, и, кроме того, напрямую связан с очень удачным критерием метода κ-средних (см. ниже в этом разделе и [Mirkin 1996]). Оказалось также, что исходный суммарный критерий не так уж плох, если предварительно вычесть из матрицы «шумовые связи». В связи с аппроксимационным подходом Миркин рассматривал вычитание из связей константы, «мягкого порогового значения» или «сдвига нуля шкалы связей» [Куперштох и др. 1976; Миркин 1985]. Сравнительно недавно был предложен так называемый критерий модулярности, в котором та же сумма используется после вычитания из каждой связи ее «случайной составляющей», пропорциональной произведению суммарных связей в соответствующих строке и столбце (см. [Newman, Girvan 2004; Newman 2006]). Вероятностная модель, основанная на случайном разрушении основной структуры разбиения, приводящая к подобному критерию, была рассмотрена в работе [Ben-Dor et al. 1999].

Еще более радикальное преобразование суммарного критерия к виду

$$as(S_1, S_2) = a(S_1, S_1)/a(S_1, I) + a(S_2, S_2)/a(S_2, I), \qquad (1.7)$$

где $a(S_f, I)$ – суммарная связь элементов S_f со всеми элементами I, была предложена в работе [Shi, Malik 2000], которые на самом деле рассматривали эквивалентную задачу минимизации дополнительного критерия, названного ими нормализованным разрезом, и которая может быть переформулирована в терминах так называемого Лапласиана $L(A)$ как задача минимизации отношения Рэлея, известного в теории собственных чисел и векторов, и тем самым решаться с помощью так называемого спектрального подхода [Von Luxburg 2007], при котором кластеры получаются путем огрубления собственных векторов Лапласовой матрицы (см. также разд. 1.4.2.3). Критерий (1.7) получил широкую популярность, поскольку оказался очень удачным для решения проблемы сегментации изображений. По нашему мнению, построение Лапласиана – это еще одна нормализация данных, очень интересная, поскольку потенциально может использоваться для серьезного уточнения понятий, связанных с моделированием данных, однако же не имеющая в данный момент надежного обоснования – в наших экспериментах применение этого преобразования иногда действительно уточняло кластерную структуру, а иногда разрушало, и неясно почему [Mirkin, Nascimento 2011].

1.3.Б Критерий метода к-средних

В отличие от данных о сходстве, удачный критерий для таблиц объект – признак был найден достаточно рано: сумма расстояний от объектов до центров соответствующих кластеров, как показано на рис. 7.

Переходя к математической формулировке, рассмотрим предобработанную матрицу объект – признак $Y = (y_{iv})$, где столбцы $v = 1,..., V$ соответствуют признакам, а строки $i \in I$ – объектам. Кластерная структура метода κ-средних задается разбиением S множества объектов на K непересекающихся кластеров, $S = \{S_1, S_2, ..., S_K\}$, представляемых таким образом через кластеры S_k, и центроиды $c_k = (c_{k1}, c_{k2}, ..., c_{kV})$, $k = 1, 2, ..., K$. Тогда минимизируемым критерием метода является сумма расстояний $d(y_i, c_k)$ от объектов y_i до соответствующих центроидов c_k:

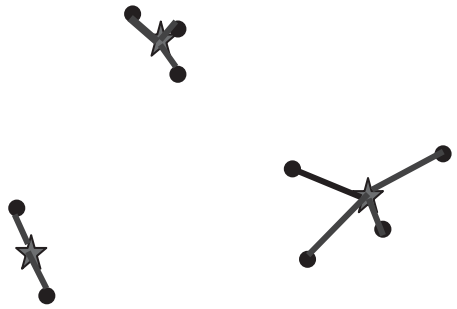

Рис. 7. Суммарные расстояния между точками, представляющими объекты, и центрами соответствующих кластеров, представленными звездами – суть критерия метода *к*-средних

$$W(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k) . \qquad (1.8)$$

В случае, когда $d(y_i, c_k)$ это квадрат Евклидова расстояния, данный критерий может быть выражен как аппроксимационный критерий наименьших квадратов в простой модели типа той, что используется в дисперсионном анализе.

Согласно этой модели, каждый объект, представленный строкой $y_i = (y_{i1}, y_{i2}, ..., y_{iV})$ матрицы Y, равен, с точностью до небольших погрешностей, центроиду соответствующего кластера c_k:

$y_{iv} = c_{kv} + e_{iv}$ для всех $i \in S_k$ и всех $v = 1, 2, ..., V$.

Сумма квадратов погрешностей этой модели, очевидно, равна

$$L^2 = \sum_{i \in I} \sum_{v \in V} e^2_{iv} = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} (y_{iv} - c_{kv})^2 ,$$

что, в свою очередь, совпадает с выражением для критерия $W(S, c)$ при расстоянии d, равном квадрату Евклидового расстояния.

Оказалось, как это типично для математико-статистического оценивания, что последний критерий возникает как частный случай метода максимального правдоподобия при оценки модели смеси Гауссовых распределений в очень ограничительной ситуации, когда ковариационные матрицы всех кластеров равны $\sigma^2 I$, где I – единичная

матрица, а σ^2 – одна и та же дисперсия, т.е. все признаки независимы и имеют одинаковое распределение Гаусса.

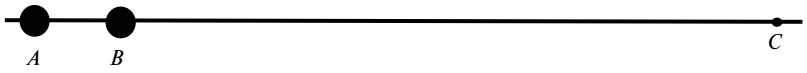

Рис. 8. Три множества точек в позициях *A*, *B* и *C* надо разбить на две группы: что лучше, *A* и *B* против *C* или *A* против *B* и *C*?

Некоторые считают, что поэтому метод к-средних применим только в данных предположениях («кластеры – сферы одного и того же радиуса»), но это не так: как известно, из того, что утверждение *M* влечет утверждение *H* вовсе не следует, что *H* влечет *M*. Ситуация здесь сходна с той, что возникает при применении линейной регрессии: линейная регрессия действительно работает для нормальных распределений, но может применяться и просто для того, чтобы аппроксимировать данные, не обязательно Гауссовые, прямой линией.

Особенность данного критерия – он «старается» увеличить равномерность распределения объектов по кластерам «любой ценой», иногда вопреки здравому смыслу. Рассмотрим пример (рис. 8).

Имеется три множества одномерных точек, два по сто точек сидят в позициях *A* и *B*, а одна точка – в позиции *C*, которая отстоит от *B* в 5 раз дальше, чем *A*, например, $AB = 2$ и $BC = 10$. Попробуем разделить 201 точку на два класса. Геометрически надо бы объединять *A* и *B*, поскольку они значительно ближе. Однако нетрудно доказать, что критерий метода *к*-средних будет меньше, если *B* объединить с *C*, поскольку численности этих двух кластеров значительно более сбалансированны. Однако подобные ситуации встречаются редко, и критерий остается наиболее популярным.

1.3.B. Эквивалентные переформулировки критерия к-средних

Mirkin (1996, 2011) дает следующие эквивалентные формулировки критерия, которые могут оказаться полезными в зависимости от того, какая цель преследуется при кластер-анализе:

(1) Максимизировать объясненную часть разброса данных

$$B(S,c) = \sum_{k=1}^{K} \sum_{v \in V} {c_{kv}}^2 N_k,$$

где N_k – численность k-го кластера, а c_k – его центроид.

Действительно, критерий $W(S, c)$ связан с $B(S, c)$ уравнением $T(Y) = B(S, c) + W(S, c)$, где $T(Y) = \sum_{v=1}^{V} \sum_{i \in I} y_{iv}^2$ – это сумма квадратов всех элементов матрицы данных, называемая (квадратичным) разбросом данных и не зависящая от кластеров. Согласно этому уравнению, $W(S, c)$ – это сумма квадратов погрешностей, характеризующая необъясненную часть разброса данных.

(2) Слегка переписанный критерий $B(S, c)$ имеет смысл как критерий поиска кластеров с центрами, наиболее удаленными от 0 (напомним, нуль заранее перенесен в точку центра тяжести).

$$B(S,c) = \sum_{k=1}^{K} \sum_{v \in V} c_{kv}^2 N_k = \sum_{k=1}^{K} N_k d(0, c_k).$$

(3) Одновременно это максимум суммарных скалярных произведений:

$$B(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} < y_i, c_k > .$$

При этом следует иметь в виду, что Евклидовы расстояния в $W(S, c)$ не зависят от положения 0, тогда как скалярные произведения в $B(S, c)$ – зависят.

Минимизировать

$$D(S) = \sum_{k=1}^{K} \sum_{i,j \in S_k} d(y_i, y_j) / N_k ,$$

где $d(y_i, y_j)$ – квадрат Евклидова расстояния между объектами i и j.

Максимизировать

$$C(S) = \sum_{k=1}^{K} \sum_{i,j \in S_k} a_{ij} / N_k,$$

где $a_{ij} = <y_i, y_j>$ – скалярное произведение вектор-строк, соответствующих объектам i and j.

(6) Максимизировать суммарный коэффициент ассоциации между искомым разбиением и заданными признаками:

$$\zeta(S) = \Sigma_{v \in V} \eta^2(S, v),$$

если все признаки – количественные; при этом η^2 – известная в статистике характеристика, так называемое корреляционное отношение. Если же, скажем, все признаки – неколичественные, то суммарный коэффициент может выражать такие меры ассоциации, как хи-квадрат Пирсона по таблице сопряженности между искомым разбиением и заданным категоризованным признаком [Mirkin 2011].

(7) Минимизировать сумму квадратов невязок в следующей модели:

$$y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + e_{iv},$$

где бинарный вектор принадлежности $z_\kappa = (z_{i\kappa})$ характеризует кластер S_k, так что z_{ik}, $= 1$ для $i \in S_k$, тогда как остальные компоненты этого вектора равны 0. Эта модель фактически совпадает с исходной аппроксимационной моделью для метода κ-средних в начале этого раздела, так как кластеры не пересекаются, т.е. только одно слагаемое в сумме не равно нулю, и это слагаемое – в точности c_k. Однако данная модель в точности соответствует уравнениям сингулярного разложения матрицы Y по сингулярным векторам и значениям, что показывает, что рассматриваемая задача кластер-анализа может рассматриваться как перенос задачи о сингулярном разложении на случай, когда накладывается ограничение, состоящее в том, что векторы z_k должны иметь 1/0 значения. Этот факт в какой-то мере лежит в основе успешности так называемого спектрального подхода к кластер-анализу: вклады отдельных кластеров в разброс данных, $\sum_{v \in V} c_{kv}^2 N_k$ аналогичны вкладам отдельных сингулярных векторов, равных, как известно, квадратам соответствующих сингулярных значений μ_k^2. Точнее, величины μ_k^2 являются собственными векторами матрицы YY^T, что выражается в терминах так называемого отношения Рэлея:

$$\mu_k^2 = z_k^T YY^T z_k / z_k^T z_k = \Sigma_v c_{kv}^2 |S_k|,$$

последнее равенство в котором выводится из условия, что вектор z_k состоит из 1/0 элементов. Следует отметить, что этот факт остается более или менее неизвестным специалистам, несмотря на неоднократные публикации автора.

Эти формулировки могли бы быть использованы в альтернативных алгоритмах k-средних, что, по-видимому, остается уделом будущего, хотя частичное освоение уже началось (см. метод аномального кластера в разделе 1.4.2.1.B).

1.4. Методы кластер-анализа

Рассмотренные выше оптимизационные задачи кластер-анализа, вообще говоря, не имеют простых решений. Поэтому в основном используются следующие методы:

– методы оптимизации общих функций – градиентный алгоритм и чередующаяся оптимизация;

– методы комбинаторной локальной оптизации, основанные на переборе в ограниченных окрестностях уже полученных решений;

– релаксационные методы, которые решают ассоциированную проблему непрерывной минимизации (релаксированная проблема), а затем огрубляют решение до комбинаторного кластерного результата – особенно часто это делается через так называемые методы положительно-полуопределенного программирования (см., например, [De Biel et al. 2006]) или собственные векторы матриц (спектральный кластер-анализ);

– алгоритмы, имитирующие природу; основаны на организации «сообществ» допустимых решений и их «эволюции», согласно модельным правилам, широко использующим случайный выбор – после некоего числа итераций перехода от поколения к поколению выбирается лучшее из так порожденных решений – очевидно, относительно случайным образом.

Мы обсудим здесь несколько подходов, связанных с основными рассмотренными кластерными структурами и типами данных. Речь пойдет прежде всего о следующих структурах:

– отдельные кластеры;

– разбиения;

– иерархии;

– бикластеры.

Рассмотрим их по порядку.

1.4.1. Выделение отдельного кластера

Идея о том, что все множество совсем не обязательно распадается на совокупность кластеров, а содержит от силы один-два кластера, оставляя другие объекты в стороне, не нова. Она была высказана с самого начала: например, таково понятие *B*-кластера (Harman, Holzinger, 1941). Необходимо упомянуть такие концепции теории графов, как компонента, бикомпонента и (максимальная) клика, формализующие понятие об отдельном кластере и, в случае компоненты и бикомпоненты, дающие эффективные алгоритмы их построения. К сожалению, в практической работе эти понятия оказываются слишком «ригидными» – в частности, исходная информация в них приведена к виду обыкновенных графов. Далее описываются следующие пять различных подходов: (1) кластеры по Апресяну [Апресян 1966], (2) движущийся кластер [Елкина, Загоруйко 1966], (3) аппроксимационный кластер [Mirkin, 1996], (4) монотонный кластер [Mullat 1976; Кузнецов, Мучник, 1981] и (5) логический таксон [Лбов, Пестунова, 1985]. Может показаться, что автор слишком увлекся отечественными авторами, но это так: зарубежные авторы не опубликовали пока ничего интересного по данному вопросу.

1.4.1.1. Кластер по Апресяну

При заданной матрице коэффициентов различия или расстояния $D = (d(i, j))$, $i, j \in I$, подмножество S называется A-кластером, если для любых различающихся $i, j, k \in I$ таких, что $i, j \in S$, а $k \notin S$, $d(i, j) \subset d(i, k)$ [Apresian 1966]. Нетрудно видеть, множество всех A-кластеров для данной матрицы D образует иерархию, т.е. если два A-кластера пересекаются, то один из них – часть второго. За истекшее время появились различные обобщения этого понятия; вероятно, самым популярным из них является понятие слабого кластера, в котором $d(i, j) \subset \max(d(i, k), d(j, k))$ [Bandelt, Dress 1989] – для A-кластеров имеет место более сильное условие $d(i, j) \subset \min(d(j, k), d(j, k))$.

1.4.1.2. ФОРЭЛЬ: движущийся кластер

При заданной стандартизованной матрице данных $Y = (y_{iv})$, где $i \in I$ – объекты, а $v \in V$ – признаки, зададимся каким-либо пороговым значением «кластерного радиуса» $D > 0$ и определим гипотетический кластер S как множество объектов, которые ближе к центру тяжести множества, чем D, $S = \{i : d(i, g) < D\}$, где $d(i, g)$ – Евклидово рассто-

яние между строкой $i \in I$ и центром тяжести g. Теперь начнем итерации: вычислим центр тяжести S, gs и переопределим S вокруг нового центра, $S = \{i : d(i, gs) < D\}$, продолжая их, пока не сойдутся. При необходимости последующие кластеры могут быть рассчитаны таким же способом на множестве, оставшемся после удаления уже кластеризованных объектов. Радиус D можно подобрать, анализируя результаты нескольких попыток. Этот алгоритм был предложен Загоруйко и Елкиной (1966) и назван ими ФОРЭЛЬ (ФОРмальный Элемент). На реальных данных этот алгоритм производит несколько крупных кластеров и множество мелких, что в прежние времена рассматривалось как недостаток, но теперь, в контексте «частичного» кластер-анализа, данная особенность – скорее преимущество.

1.4.1.3. Аппроксимационный кластер

При заданной матрице связи между объектами $B = (b_{ij})$, искомый кластер S представляется бинарным вектором принадлежности $s = (s_i)$, где $s_i = 1$, если $i \in S$, и $s_i = 0$ – в противном случае, а также положительным (в норме) уровнем интенсивности λ, так, чтобы суммарная разница квадратов $L(S) = \Sigma i_{,j \in I}(b_{ij} - \lambda s_i s_j)^2$ была минимальной (Миркин 1985). Нетрудно видеть, что при заданном S оптимальное значение λ равно средней связи $b(S)$ внутри кластера S, а минимум суммы квадратов достигается при максимизации суммы внутренних связей в S, деленной на $|S|$ – численность S (т.е. одно слагаемое суммарного критерия Дорофеюка – Бравермана (1.6)), что также равно произведению $b(S)|S|$ – компромиссный критерий, максимизация которого требует одновременной максимизации двух противоречивых критериев: средней связи (убывает при увеличении кластера) и численности кластера.

Соответствующий метод локальной оптимизации, в двух версиях, при фиксированной или оптимизируемой интенсивности λ, обозначаемых ADDI и ADDI-S, был описан Миркиным в различных публикациях, включая [Mirkin 1987, 1996]. Суть метода: при текущем S и произвольном объекте i, проверяется, как изменится критерий метода при добавлении i к S, если i не принадлежит S, или при вычитании i из S, если $i \in S$. Оказывается, это изменение линейно зависит от разности $Dif\,(i) = b\,(i, S) - \lambda/2$, где b (i, S) – средняя связь между i и S. Таким образом, разности Dif (i) вычисляются для всех i, после чего тот объект i^*, для которого уменьшение критерия макси-

мально, переводится в S или удаляется из S. Если же $Dif(i^*)$ положительно, то процесс закончен, кластер S – финальный.

К преимуществам метода относятся: (а) математически доказанное свойство «тесноты» получаемого S: для всякого $i \in S$ средняя связь с S превышает $\lambda/2$, а для всякого $i \notin S$ средняя связь с S меньше, чем $\lambda/2$ [Mirkin 1976]; (б) декомпозиция разброса связей, т.е. суммы их квадратов, на части, объясненные и необъясненные кластерной структурой; (в) возможность получения пересекающихся решений. Часть ADDI, связанная с использованием «мягкого» порогового значения $\pi = \lambda/2$, позже была описана (без ссылок) в работе [Ben-Dor et al. 1999] и названа алгоритмом CAST; этот метод получил популярность в биоинформатике.

1.4.1.4. Монотонный кластер

Оригинальный подход, основанный на оценке связей между объектами и множествами объектов, был предложении Муллатом (1976) и адаптирован для анализа организационных и других систем Мучником и его соавторами (из многих публикаций отметим только [Кузнецов, Мучник 1981; Aaremaa 1985]). Рассмотрим монотонную функцию сходства $f(i, S)$ между всеми объектами $i \in I$ и подмножествами $S \subset I$ как исходную информацию. Монотонность подразумевает, что (а) $f(i, S) \le f(i, S \cup T)$ для всех S и T или (б) $f(i, S) \ge f(i, S \cup T)$ для всех S и T. В типичной ситуации такую функцию легко можно определить на основе данных любой природы, включая оцифрованные текст или изображение. Например, при заданной неотрицательной матрице связи (или, что то же самое, взвешенном графе) $A = (a_{ij})$ можно определить $f(i, S)$ как $\Sigma_{i \in S} a_{ij}$ или $\min_{i \in S} a_{ij}$ так что f монотонна по S в нужную сторону. Рассмотрим, для определенности, монотонно возрастающую $f(i, S)$ и определим функцию множеств $F(S) = \min_{i \in S} f(i, S)$, характеризующую «наислабейшее звено» в S. Оказывается, такие функции $F(S)$ можно максимизировать «жадным» образом, подбирая на каждом шаге наилучшего кандидата, что ведет к «оптимальной» последовательности объектов, которая определяет не только «ядро» – множество S, максимизирующее $F(S)$ как фрагмент этой последовательности, но и совокупность его «оболочек» – объемлющих фрагментов. Функции «наислабейшего звена», оказывается, – те и только те, которые удовлетворяют так называемому условию квазивыпуклости

$$F(S \cup T) \ge \min(F(S), F(T)) \qquad (1.12)$$

для всех S, $T \subset I$ (результат А.В. Малишевского, опубликованный в его посмертном сборнике [Малишевский 1998, с. 171–172]. Индикатор подмножества $T \subset I$, функция $G(S)$, определенная правилом: $G(S) = 0$ для всех S, не совпадающих с T, и $G(T) = 1$, также квазивыпукла. Однако проблема максимизации индикаторной функции, как известно, NP-полна, что противоречит возможности ее «жадной» максимизации, на первый взгляд. На самом деле это не так, поскольку соответствующая функция связи $g(i, S)$ дает «оракульную» информацию о G, недоступную в классической формулировке задачи.

Функции «наислабейшего звена» связаны с порядковыми структурами типа конечных «выпуклых геометрий» (см. [Mirkin, Muchnik 2002]). С другой стороны, это понятие может быть адекватно использовано для описания особенностей организационных и интернетовских систем (см. [Кузнецов, Мучник 1981; Мгеладзе, Гоциридзе 2009]). Скорее всего, этому подходу еще предстоит сыграть свою роль в анализе структуры сложных систем.

1.4.1.5. Логический таксон

Любой предикат, сформированный из признаков, как, например, «$y_1 = 3$ и $y_2/y_3 > 5$», определяет множество объектов S, удовлетворяющих ему. Интересно, что в этом контексте никаких ограничений на типы шкал не накладывается! Доля f множества S в I – это наблюденная частота истинности предиката. С другой стороны, каждая составляющая предиката, связанная с отдельным признаком, в нашем примере «$y_1 = 3$» и «$y_2/y_3 > 5$» имеет свою частоту $f1$ и $f2$. Эти индивидуальные частоты легко скомбинировать так, чтобы определить ожидаемую частоту предиката ef по правилам вероятности для логических операций – в нашем примере $ef = f1*f2$. Чем больше разница между f и ef, тем лучше, неожиданнее S. Этот критерий, вместе с «жадным» алгоритмом его оптимизации, был предложен Лбовым и Пестуновой (1985) на основе более ранних публикаций Г.С. Лбова. Аналогичный упрощенный критерий, отношение $P(y_1, y_2, ..., y_n) / (P(y_1) P(y_2) ... P(y_n))$, позже использовался основателями компании Megaputer Intelligence, одной из очень немногих успешных международных компаний, выдвинутых российским научным сообществом после распада СССР (см. [Kiselev et al. 1999]).

1.4.2. Построение разбиений

Разбиение – самая популярная кластерная структура, и число подходов здесь исчисляется десятками, если не сотнями. Мы отдельно рассмотрим наиболее популярные подходы для данных вида объект – признак и для данных вида матриц связи.

1.4.2.1. Метод *к*-средних

K-средних – самый популярный метод построения разбиения по матрице объект – признак, включен во все пакеты статистики и разработки данных, такие как SPSS, SAS, Clementine, iDA tool и DBMiner.

В методе *к*-средних процесс порождения разбиений с центроидами начинается с некоторых начальных точек – центроидов (экспериментально показано, что лучше всего, если это объекты анализируемого множества, а не абстрактные местоположения), а затем осуществляется последовательность итераций. Каждая итерация состоит из двух шагов:

1. Обновление кластеров: При заданных *K* центроидах c_k ($k = 1,2,...,K$), каждый объект $i \in I$ приписывается к одному из центроидов согласно так называемому правилу минимального расстояния: вычисляются расстояния от i до каждого c_k, после чего i отходит к ближайшему из c_k, (рис. 9(б)). Объекты, приписанные к c_k, образуют обновленный кластер S_k ($k = 1, 2, ..., K$), (рис. 9(c)).

2. Обновление центроидов: Для каждого кластера S_k вычисляется его центр тяжести (внутрикластерное среднее по каждому признаку), который и объявляется новым центроидом c_k' ($k = 1, 2, ..., K$) (рис. 9(d)).

В результате процесс приходит к тому, с чего начинался, система центроидов – проводится новая итерация, и т.д., до сходимости. В ситуации, когда в качестве расстояния используется квадрат Евклидовой метрики, данный процесс на самом деле реализует метод чередующейся минимизации критерия (1.8) этого метода относительно двух групп переменных, кластеров и центроидов, и поэтому обязательно сходится, из-за конечности общего числа разбиений на конечном множестве, хотя, конечно, получаемое решение может быть весьма далеким от оптимума как по разбиению, так и по значению критерия.

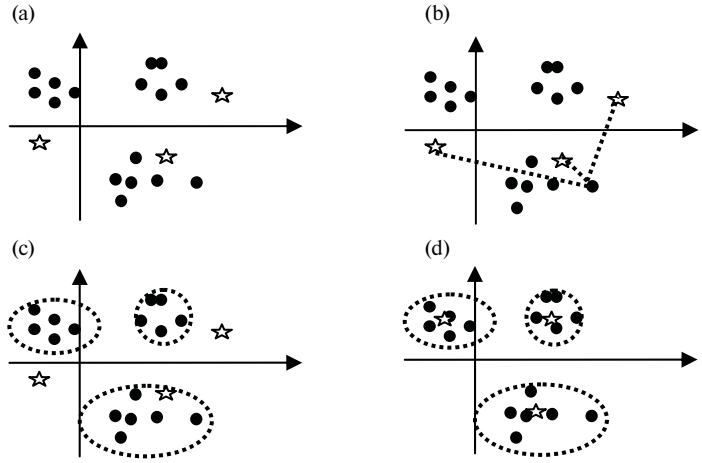

Рис. 9. Основные шаги параллельного метода *к*-средних: (a) инициализация центроидов, (b) обновление кластеров (пунктиром показаны расстояния от объекта к центроидам), (c) окончание процесса обновления кластеров, (d) окончание процесса обновления центроидов.

Процесс метода *к*-средних обновляет центроиды так, чтобы они попали в места относительно высокой плотности объектов. С другой стороны, его можно интерпретировать как процесс формирования типологии, где центроиды служат моделями «типичных представителей». Этот метод работает быстро и может быть организован с экономией памяти, а также в распределенном режиме.

Имеется естественная модификация метода на ситуацию адаптивного типа, когда объекты поступают по одному, а не все сразу (режим «онлайн»), (см. [MacQueen 1967]). При этом правило минимального расстояния остается справедливым для отнесения вновь прибывшего объекта к кластеру с ближайшим центроидом. Центроиды обновляются после каждого такого присоединения: если y_i добавлен кластеру S_k численности N_k, его центроид c_k переводится в c'_k по формуле, очевидной по определению среднего:

$$c'_k = N_k c_k/(N_k + 1) + y_i/(N_k + 1).$$

Вместе с тем метод имеет определенные недостатки, связанные с отсутствием указаний о выборе числа K и начального положения центроидов, их неправильный выбор может приводить к плохим резуль-

татам (рис. 10). Необходимы средства выявления значимых признаков. Однако алгоритм не способен отделить кластерообразующие признаки от шумовых, что может приводить к нежелательным результатам. Ниже будут предложены некоторые способы решения указанных проблем.

Рис. 10. Пример недостатка метода κ-средних – зависимость кластеров от начальных центроидов, представленных звездочками: справа и слева показаны кластеры, возникающие при соответствующей инициализации. Правильные кластеры – те, что справа

1.4.2.2 Модификации метода κ-средних

1.4.2.2. А. Имитирующие природу алгоритмы

Поскольку итерации метода κ-средних не очень далеко уводят от исходных центроидов, возникает идея использования более мощных алгоритмов, которые бы выбивали процесс из локального оптимума. В этом плане перспективным считаются так называемые подходы, инспирированные природой. В отличие от классических методов оптимизации, основанных на «доводке» единственного решения, эти подходы обрабатывают целую «популяцию», состоящую из многих решений одновременно. При этом основное внимание обращается на процесс перехода от данной популяции к популяции следующего поколения. Алгоритм, инспирированный природой, начинает с произвольной популяции и генерирует последовательность ее поколений, каждый раз регистрируя наилучшие решения. После определенного числа итераций алгоритм заканчивает работу, и наилучшее из найденных решений выдается в качестве результата. С точки зрения метода κ-средних основное значение имеет способ представления кластерной структуры в виде единственной последовательности чисел – строки. В литературе рассматриваются оба элемента кластерной структуры – (а) разбиение и (б) центроиды как основа кодирования решения в виде строки. Каждый из этих элементов позволяет легко восстановить второй согласно основной процедуре метода κ-средних.

Чтобы представить разбиение единой строкой, фиксируется определенное упорядочение объектов, после чего формируется последовательность номеров кластеров, соответствующая этому упорядочению. Если, например, имеется восемь объектов, упорядоченных как $e1$, e 2, e 3, e 4, e 5, e 6, e 7, e 8, то строка 12333112 представляет разбиение S на три класса: $S1 = \{e\ 1, e\ 6, e\ 7\}$, $S2 = \{e\ 2, e\ 8\}$, and $S3 = \{e\ 3, e\ 4, e\ 5\}$, что очевидно, если сопоставить номера и объекты:

e 1 e 2 e 3 e 4 e 5 e 6 e 7 e 8
1 2 3 3 3 1 1 2

При этом допустимы только такие кодирующие строки, в которых не пропущен ни один из номеров, так как, если, скажем, номер 2 пропущен, то класс S – пустой, что недопустимо при заданном K.

Чтобы представить центроиды в виде строки, надо зафиксировать порядок кластеров, после чего сформировать вектор, состоящий из центроидов в данном порядке. Если, скажем, объекты из вышеприведенного примера характеризуются пятью признаками, то три 5-мерных центроида образуют вектор-строку из 15 чисел – координат этих центроидов.

Ниже будут рассмотрены три наиболее популярных типа алгоритмов, имитирующих природу:

– генетический алгоритм;

– эволюционный алгоритм;

– алгоритм роя частиц.

В генетических алгоритмах строка, представляющая решение, называется хромосомой. Рассмотрим для удобства случай, когда строки представляют разбиения. Вычисления начинаются со случайной генерации определенного числа, p, хромосом, каждая из которых оценивается затем критерием метода. Наилучшая из этих хромосом называется элитой и запоминается. Затем производится формирование пар хромосом для выведения потомства, обычно случайным образом, причем вероятности могут отражать значение критерия, т.е. «приспособленность» хромосом (см. [Murthy, Chowdhury 1996; Yi Lu et al. 2004]). Потомство производится путем кроссовера, состоящего в том, что в каждой паре обе хромосомы разрываются в одном и том же, случайно выбранном, месте, после чего начальный фрагмент одной хромосомы соединяется с конечным фрагментом другой. Это потомство затем подвергается мутации – с некоторой небольшой вероятностью номера кластеров меняются на другие, после чего процесс

порождения нового поколения почти завершен. Осталось только провести поддержку элиты. Для этого, например, лучшая (по критерию метода κ-средних) из полученных хромосом сравнивается с элитной и заменяет ее, если она лучше. В противном случае элитная хромосома вставляется в новое поколение вместо самой худшей хромосомы. Y. Lu et al. (2004) отмечают, что генетический алгоритм приводит к значительно лучшим результатам, если после мутации все кластеры обновляются в соответствии с Правилом минимального расстояния.

С вычислительной точки зрения использование разбиения в качестве хромосомы имеет недостаток – ее длина определяется количеством объектов, которое может быть очень велико. Этот недостаток преодолевается, если использовать строки, представляющие центроиды, длина которых не зависит от числа объектов. Более того, использование центроидов позволяет использовать свойства непрерывности вещественных чисел и вместо мутации делать небольшие изменения элементов центроидов. Это, естественно, приводит к так называемым эволюционным алгоритмам.

В эволюционном алгоритме мутации происходят следующим образом (см. [Bandyopadhyay, Maulik 2002]). Обозначим $minW$ и $maxW$ минимальное и максимальное значения критерия на элементах популяции. Определим радиус хромосомы R как долю максимума $maxW$, достигнутого в ней: $R = (W\text{-}minW)/(maxW\text{-}minW)$. После этого случайно генерируется интенсивность мутации δ в интервале от $-R$ до R.

Обозначив $minv$ и $maxv$ минимум и максимум признака v на всем множестве ($v = 1,..., V$), v компонента xv центроида c_k в хромосоме меняется по правилу

$xv+\delta*(maxv - xv)$, если $\delta \geq 0$ или

$xv+\delta*(xv - minv)$ – в противном случае.

Заметим, что наилучшая хромосома, в которой $W = minW$, в этом процессе не меняется, так как для нее $R = 0$.

Более продвинутая схема эволюции, названная «дифференциальной эволюцией», рассмотрена в работе [Paterlini, Krink 2006].

Несколько другой биологический процесс, отыскания пищи роем пчел, имитируется в методе роя частиц. Рой движется в случайном направлении, при этом запоминая наилучшие места из посещенных. Каждая частица роя характеризуется следующими атрибутами:

– вектор состояния x – KV-мерная строка вектора центроидов;
– оценка приспособленности $f(x)$ критерием метода;
– вектор скорости z той же размерности, что и x;
– запись наилучшего из уже посещенных состояний b.

Наилучшая позиция всего роя bg определяется как наилучшее из индивидуальных состояний b.

На итерации t ($t = 0, 1, ...$) следующее состояние определяется добавлением вектора скорости к вектору состояния:

$$x(t + 1) = x(t) + z(t + 1),$$

где $z(t + 1)$ определяется текущей скоростью и наилучшими состояниями:

$$z(t+1) = z(t) + \alpha\,(b - x(t)) + \beta\,(bg - x(t)),$$

где α и β – равномерно распределенные случайные величины, обычно из интервала между 0 и 2; слагаемое $\alpha(b - x(t))$ характеризует «когнитивную», а $\beta(bg - x(t))$ «социальную» компоненту процесса.

Начальные значения $x(0)$ и $z(0)$ генерируются случайно внутри заранее определенной области допустимых значений.

Следует отметить, что методы, имитирующие природу, требуют довольно много вычислительного времени.

1.4.2.2.Б. Нечеткие кластеры

Нечеткий кластер задается функцией принадлежности $z = (z_i)$, $i \in I$, так что величины принадлежности z_i ($0 \subset z_i \subset 1$) интерпретируются как степени принадлежности объектов i кластеру (для четких кластеров значения z_i могут быть только 1 или 0).

Нечеткое разбиение с системой центроидов – это K центроидов $c_k = (c_{k1}, ..., c_{kv}, ..., c_{kV})$ в пространстве признаков и векторов принадлежности $z_k = (z_{1k}, ..., z_{ik}, ..., z_{Nk})$, $0 \subset z_{ik} \subset 1$, таких что $\Sigma_k z_{ik} = 1$ для всех $i \in I$: т.е. полная принадлежность как бы распределена между кластерами.

Билинейная модель сингулярного разложения матриц в разд. 1.1.3 для этого случая приводит к довольно необычным структурам, в которых центроиды стремятся занять место крайних точек облака объектов, что может иметь свою ценность, так как эта структура устойчива, когда много «средних», не очень выделяющихся объектов добавляется к данным – см. [Nascimento 2005]. Очень популярно другое обобщение критерия κ-средних:

$$F(\{c_k, z_k\}) = \sum_{k=1}^{K} \sum_{i=1}^{N} {z_{ik}}^{\alpha} d(y_i, c_k)$$

где d – квадрат Евклидова расстояния, а α – показатель, влияющий на степень «контрастности» оптимальных значений принадлежности: при $\alpha = 1$, оптимальные принадлежности равны 0 или 1, т.е. не являются нечеткими. На практике обычно полагают $\alpha = 2$.

Хотя глобально оптимизировать этот критерий нелегко, метод чередующейся минимизации приводит к замкнутым формулам и широко используется в литературе как нечеткая версия κ-средних, иногда называемая методом c-средних. Каждая итерация состоит из двух шагов:

(1) при заданных центроидах кластерные принадлежности рассчитываются по формуле

$$z_{ik} = 1 / \sum_{k'=1}^{K} [d(y_i, c_k) / d(y_i, c_{k'})]^{\frac{1}{\alpha-1}};$$

(2) при заданных принадлежностях центроиды рассчитываются по формуле

$$c_{kv} = \sum_{i=1}^{N} {z_{ik}}^{\alpha} y_i / \sum_{i'=1}^{N} {z_{i'k}}^{\alpha}.$$

Эти формулы легко получаются из необходимых условий оптимальности критерия; сходимость к локальному оптимуму гарантирована (см. [Bezdek et al. 1999]). Смысл критерия: $F = \Sigma_i F(i)$, где $F(i)$ выражает гармоническое среднее принадлежностей объекта к кластерам при $\alpha = 2$ как объяснено в [Stanforth, Mirkin, Kolossov 2007].

1.4.2.2.B. Аномальные кластеры

Этот метод – применение стратегии последовательного исчерпывания данных, в данном случае, по одному кластеру. Матричная модель (1.11) из раздела 1.3.B для случая, когда ищется всего один кластер, имеет вид

$$y_{iv} = \begin{cases} c_v + e_v, i \in S \\ 0 \quad + e_v, i \notin S \end{cases},$$

где S – искомый кластер, а c его центроид. Минимизация суммы квадратов по неизвестным S и c эквивалентна максимизации величины вклада в суммарный разброс данных

$$\mu^2 = z^T YY^T z / z^T z = c_v^{\,2}|S| = d(0, c)|S|;$$

$$W(S,c) = \sum_{i \in S} d(y_i,c) + \sum_{i \notin S} d(y_i,0);$$

или, что то же самое, минимизации однокластерной версии критерия *к*-средних, в котором центр «кластера "не *S*"» равен 0 и не может измениться, так что искомый кластер должен отстоять как можно дальше от 0. В предположении, что 0 представляет «норму» как, например, центр тяжести всего множества точек, оптимальный кластер должен быть наиболее аномальным, что объясняет его название. Несмотря на упрощенный характер модели, метод аномального кластера хорошо зарекомендовал себя в экспериментах на сгенерированных (см. [Chiang, Mirkin 2010]) и реальных данных.

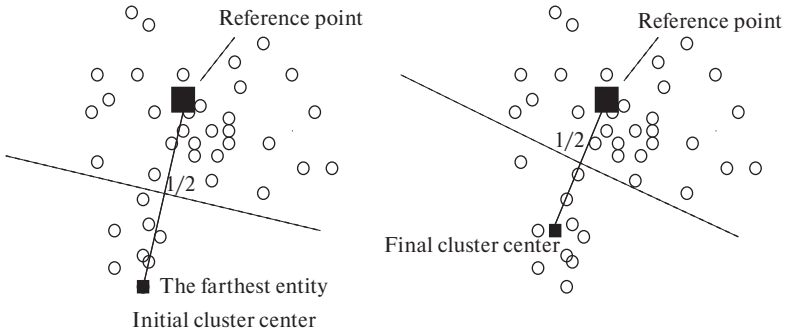

Рис. 11. Начало и конец работы метода аномального кластера (из [Mirkin 2005])

Метод аномального кластера

1. *Предварительная обработка.* Фиксируется «реперная» точка $a = (a_1, ..., a_v)$ (например, в центре тяжести «облака» объектов), которая затем переводится в 0 вычитанием ее из всех точек, представляющих объекты.

2. *Инициализация.* Точка, наиболее удаленная от 0, берется в качестве начального центроида.

3. *Обновление кластера.* Объект приписывается к центроиду *c*, если и только если $d(y_i, c) < d(y_i, 0)$; *S* образуется из всех таких объектов.

4. *Обновление центроида.* Вычисляется центр тяжести кластера *S*, *c'*, который сравнивается затем с предыдущим центроидом *c*. Если

$c' \neq c$, шаг 3 выполняется снова после того, как c заменяется на c'. В противном случае выдается решение – кластер S, его центроид c, а также значения вкладов в разброс данных – они необходимы для принятия решения о продолжении процесса кластер-анализа.

Процесс может быть остановлен, если следующий вклад слишком незначителен. В этом случае кластеры не будут покрывать все множество объектов, что может оказаться полезным, когда данные значительно зашумлены. Версия метода *к*-средних, в которой число кластеров и их положение определяются на основе итеративного применения метода аномального кластера, назван интеллектуальным, *ик*-средних (см. [Mirkin 2005]).

1.4.2.2.Г Веса для переменных

Хотя метод *ик*-средних довольно успешно справляется с преодолением недостатков метода *к*-средних, связанных с выбором K и инициализации, самый главный недостаток метода – возможно, большинства методов кластер-анализа – остается. Этот недостаток связан с тем, что метод абсолютно не защищен от наличия «шумовых» признаков, не связанных с кластерной структурой, даже если она существует (см. [Milligan 1980]). Для борьбы с этим в течение последних десятилетий были приложены определенные усилия, которые привели к модификации критерия метода, похожей на ту, что используется в методе *c*-средних нечеткого кластер-анализа – только веса здесь отражают не объекты, а признаки:

$$W(S,b,c) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} b_{kv}^{\beta} d(i,c_k),$$

где b_{kv} – вес признака v в кластере k, а β– показатель влияния, от которого зависит качество работы метода. Оказалось, что чередующаяся минимизация этого критерия по указанным трем группам признаков, как и в нечетких *к*-средних, приводит к замкнутым формулам, лежащим в основе трехшаговых итераций (см. [Huang et al. 2005]). Согласно этим формулам, чем больше разброс признака в кластере, тем меньше должен быть вес этого признака, так что шумовые признаки хорошо выявляются этим методом. Недостаток метода – некоторая искусственность показателя степени β, которую удалось снять в работе [Amorim, Mirkin 2011] за счет перехода от квадрата Евклидова расстояния к β-й степени расстояния Минковского d_{β}, опреде-

ляемой формулой $d_\beta(x, y) = \Sigma_v |x_v - y_v|^\beta$. При таком определении расстояния коэффициент в $b_{kv}{}^\beta$ в критерии вносится внутрь формулы расстояния, и вес признака превращается в масштабирующий коэффициент, так что критерий возвращается к стандартному формату критерия (1.8) для метода *к*-средних. Модификация Минковского, *мк*-метод, оказалась лучше исходного метода как в традиционной, так и «интеллектуальной» версиях. Правда, остается непонятным – как подбирать значение β – от него сильно зависят результаты, а его наилучшее значение сильно зависит от структуры данных (см. [Amorim, Mirkin 2011]).

1.4.2.2.Д. EM-метод для смеси распределений

Это направление – так называемый «кластер-анализ, основанный на модели», поскольку здесь, в отличие от остальных методов есть модель порождения данных. Согласно модели смеси распределений, наблюденные данные – это независимая случайная выборка из распределения, функция плотности которого имеет вид

$$f(x \mid \theta) = \sum_{k=1}^{K} \tau_k f_k(x \mid \theta_k),$$

где отдельные кластеры представляются одномодальными функциями плотности $f_k(x|\theta_k)$, а τ_k – их положительные вероятности ($k = 1, ..., K$). При заданной выборке $X = \{x_i\}$ оценки максимального правдоподобия параметров $\theta = \{\theta_k\}$ определяются путем максимизации логарифма величины максимального правдоподобия, представляемой обычно в виде $L(g, l) = \Sigma_i log f(x_i|\theta) = \Sigma_{ik} g_{ik}(l_{ik} - log\, g_{ik})$, где $g_{ik} = \tau_k f(x_i|\theta_k) / \Sigma_k p_k f(x_i|\theta_k)$ – так называемые постериорные вероятности принадлежности объекта *i* кластеру *k*, а $l_{ik} = log(\tau_k f(x_i|\theta_k))$. Представление логарифма функции правдоподобия в виде $L(g, l)$ было впервые сделано в работе [Rubin et al. 1977], в которой оно использовано для чередующейся максимизации этой функции (итерации, организованные чередующейся оптимизацией отдельно по каждой из указанных двух групп переменных) – этот алгоритм назван «EM-алгоритм» (Expectation-Maximization); он или его модификации используются практически во всех разработках, основанных на модели распределения. Применительно к смеси распределений EM-алгоритм может быть представлен следующим образом (см. [Biernacki et al. 1990]).

Данные рассматриваются в виде $x_i = (y_i, z_i)$, где $z_{ik} = 1$, если объект i из κ-го кластера, или = 0 – в противном случае. В предположении, что все z_i независимо и одинаково распределены согласно мультиномиальному закону с вероятностями $\tau_1,\ldots,\tau_K$, логарифм правдоподобия имеет вид

$$l(\theta,\tau,z \mid x) = \sum_{i\in I}\sum_{k=1}^{K} z_{ik} \log[\tau_k f_k(y_i \mid \theta_k)] .$$

Е-шаг вычисляет «оптимальное» значение величин z_{ik} в предположении, что все остальные аргументы функции L известны. Это делается путем вычисления Байесовской вероятности

$$p_k = \frac{\tau_k f_k(y_i \mid \theta_k)}{\sum_{l=1}^{K} \tau_l f_l(y_i \mid \theta_l)}$$

и помещения i в тот класс k, для которого p_k максимум. На M-шаге определяются оптимальные значения θ и τ внутри каждого кластера. Например, для Гауссового распределения нужно просто посчитать эмпирические величины – долю кластера в общем объеме данных (оценка вероятности τ_k), средние внутри кластеров (оценка математических ожиданий) и эмпирические ковариационные матрицы. Несмотря на внешнюю простоту и принципиальную возможность работы с пересекающимися кластерами, использование метода затруднено из-за: (а) слишком большого количества параметров, от инициализации которых сильно зависит результат, (б) невозможности оценки ковариационной матрицы в кластерах с малым количеством объектов или (в) когда переменные, как это часто бывает, мультиколлинеарны, и (г) абсолютной неприменимости в ситуации ошибочного значения K.

В последнее время, правда, фокус внимания в этом методе перемещается на использование не Гауссовых распределений, а распределений Дирихле, которые значительно проще (например, их ковариационные матрицы имеют ранг 1, а область определения ограничена (см., например, [Bouguila, Ziou 2007]).

1.4.2.3. Разбиения по матрицам связи

При заданной матрице сходства $A = (a_{ij})$ между объектами множества I попробуем разбить I на две части, S_1 и S_2, таким образом, чтобы сходство между S_1 и S_2 было минимальным, тогда как сходство

внутри – максимальным; обзор критериев такого рода был дан в разделе 1.3.А. Перенос на случай большего числа кластеров очевиден. Методы получения оптимальных разбиений бывают следующие: (а) иерархические, (б) спектральные, (в) локальные и (г) графотеоретические. Рассмотрим их поочередно.

(а) *Иерархические алгоритмы* работают либо «агломеративно», склеивая на каждом шаге два наиболее близких кластера, начиная с тривиального разбиения на одноэлементные кластеры, либо «дивизимно» – разбивая на каждом шаге какой-либо кластер, начиная с универсального кластера, состоящего из всех объектов.

Международная мысль в области агломеративных алгоритмов основана на идее, что результаты алгоритма агломерации зависят от формулы пересчета расстояний до вновь построенного кластера. Общая формула Ланса и Вильямса (1967) относится к случаю, когда связи (расстояния) до объединенного кластера определяются, исходя из связей (расстояний) с объединившимися частями. Хорошо известный пример – метод ближайшего (дальнейшего) соседа, когда связь с объединенного кластера рассчитывается как максимум (минимум) связей с объединившимися кластерами. Эти два метода приводят к кластерам, которые в какой-то мере аналогичны графотеоретическим концепциям *компоненты связности* (ближайший сосед) и *клики* (дальнейший сосед). Hartigan (1975) доказал, что метод ближайшего соседа на самом деле решает следующую задачу: для заданной матрицы связи A найти такую ультраметрическую матрицу B, что B – ближайшая к A по сумме модулей разностей между соответствующими элементами при условии, что $B \subset A$. Матрица B называется ультраметрической тогда и только тогда, когда для любых $i, j, k \in I$ выполняется следующее неравенство: $b_{ij} \geqslant \min (b_{ik}, b_{jk})$. Хорошо известно, что ультраметрические матрицы характеризуют связи, возникающие между листьями так называемого индексированного дерева. Это корневое дерево, каждый узел которого помечен значением индекса, равным 0 для корня и 100 (или любому другому числу), приписанному листьям так, что сохраняется монотонность, т.е. индекс всякой вершины больше, чем индекс ее родителя. Связь между листьями индексированного дерева – это значение индекса, приписанное их самому позднему (наименьшему) общему предку. Нетрудно доказать, что матрица связи является ультраметрической тогда и

только тогда, когда она порождена индексированным деревом. Эти построения очевидным образом переводятся на язык расстояний.

Оказывается, между методом ближнего соседа и теорией графов существует прямая связь, основанная на понятии максимального покрывающего дерева (остова). Для данной матрицы A рассмотрим взвешенный граф с множеством вершин I, ребра которого взвешены значениями из A. Остов (покрывающее дерево) этого графа – любой связный подграф с полным множеством вершин без циклов. Длина этого дерева – сумма весов ребер. Задача об отыскании покрывающего дерева максимальной длины решается жадным алгоритмом точно – известны два алгоритма, путем отбора вершин (Прим) и путем отбора ребер (Крускал), описание их можно найти в учебниках. Оказывается, что фрагменты такого дерева, получаемые разрезом (убиранием) какого-либо ребра, – это кластеры по методу ближнего соседа и только они. Это означает, что максимальное покрывающее дерево можно получать дивизимным алгоритмом, последовательно разрезая максимальное покрывающее дерево по самым слабым ребрам.

(б) *Спектральный подход.* Рассмотрим дивизимные алгоритмы, основанные на критериях, допускающих спектральный подход (см. раздел 1.3.Г). Рассмотрим для примера суммарный критерий для двух кластеров из раздела 1.3.Г, для матрицы связей, из которой предварительно были вычтены пороговые или случайные связи. Определим N-мерный вектор $z = (z_i)$ такой, что $z_i = 1$ для $i \in S_1$ и $z_i = -1$ для $i \in S_2$. Очевидно, $z_i^2 = 1$ для любого $i \in I$, так что $z^T z = N$ – не зависит от разбиения. С другой стороны, $z^T A z = a(S_1, S_1) + a(S_2, S_2) - 2\,a(S_1, S_2) = 2(a(S_1, S_1) + a(S_2, S_2)) - a(I)$, т.е. критерий суммы внутренних связей достигает максимума тогда и только тогда, когда $z^T A z$ – максимум, так что проблема отыскания оптимального разбиения эквивалентна проблеме максимизации отношения Рэлея

$$g(z) = \frac{z^T W z}{z^T z} \tag{1.13}$$

относительно неизвестного N-мерного z с 1/–1 компонентами, где матрица W – результат предварительного преобразования матрицы A. Как известно, максимум этого отношения относительно произвольного z равен максимальному собственному значению W и достигается на соответствующем собственном векторе. Отсюда выте-

кает принцип спектрального кластер-анализа: найти этот собственный вектор, после чего определить кластеры в соответствии со знаком компонент – индексы положительных элементов – в один класс, а индексы отрицательных элементов – в другой. Иными словами, будем аппроксимировать этот собственный вектор вектором из 1 и −1, заменяя положительные компоненты на 1, а отрицательные – на −1, после чего определим S_1 как множество объектов, соответствующих 1, а S_2 – соответствующих −1.

Для критерия нормализованного разреза ситуация несколько сложнее (Shi, Malik 2000). Чтобы использовать отношение Рэлея, надо применить преобразование Лапласа.

Обозначим $w_{i+} = \sum_{j \in I} w_{ij}$ ($i \in I$) суммы строк матрицы W и определим диагональную матрицу D – все ее элементы нули, а диагональный элемент (i, i) равен w_{i+}. Тогда (нормализованный) Лапласиан определяется как $L = E - D^{-1/2}WD^{-1/2}$, где E – единичная диагональная матрица, а $D^{-1/2}$ – диагональная матрица, у которой элемент (i, i) равен $1/\sqrt{w_{i+}}$. То есть в матрице L элемент (i, j) равен $\delta_{ij} - w_{ij}/\sqrt{w_{i+}w_{j+}}$ где $\delta_{ij} = 1$, если $i = j$ или 0 – в противном случае. Оказывается, $Lf_0 = 0$ для $f_0 = (\sqrt{w_{i+}}) = D^{1/2}1_N$, где 1_N – вектор все элементы которого равны 1, т.е. 0 – собственное значение L, соответствующее собственному вектору f_0. Поскольку матрица L положительно полуопределена, 0 оказывается минимальным собственным значением. Оказывается, критерий нормализованного разреза состоит в том, чтобы найти такое разбиение $S = \{S_1, S_2\}$ на I, и, значит, соответствующий вектор s у которого $s_i = \sqrt{w_{i+}w(S_2)/w(S_1)}$ для $i \in S_1$ и $s_i = -\sqrt{w_{i+}w(S_1)/w(S_2)}$ для $i \in S_2$, такой, что отношение Рэлея для s минимально по всевозможным разбиениям на два класса. При этом векторы s и f_0 ортогональны.

Значит, можно применять спектральный подход применительно к наименьшему ненулевому собственному числу матрицы L для решения задачи о нормализованном разрезе [Shi, Malik 2000]. Или, что то же самое, к максимальному собственному числу псевдообратной матрицы L^+. Последнее связывает спектральный кластер-анализ

с так называемыми аддитивными кластерами, как описано в [Mirkin, Nascimento 2011].

В последнее время математики начали анализировать задачу аппроксимации произвольной матрицы связи с помощью ультраметрической матрицы – эта задача переборная, но есть надежда, что вероятностные алгоритмы могут принести успех; обзор и последние результаты можно найти в [Guo et al. 2010].

(в) *Локальный алгоритм* на каждом шаге рассматривает некоторое разбиение $\{S_1, S_2\}$ и производит локальное его преобразование в сторону улучшения значения критериев. Из всех возможных локальных преобразований рассматривался только перенос одного объекта из класса в класс. Для этого отыскивается объект, на котором приращение критерия максимизируется, и если оно положительно – перенос производится. Если нет – разбиение объявляется окончательным результатом. Метод работает хорошо, но на относительно небольших данных.

(г) *Графотеоретический подход*

Простейшее понятие здесь – так называемые пороговые графы: представим данную матрицу связи взвешенным графом, а потом выкинем все ребра, которые меньше некоторого заранее выбранного порогового значения. Никто не анализировал с теоретической точки зрения работу на оставшемся взвешенном графе – хотя на самом деле сегментация изображений осуществляется именно так: все расстояния, которые больше порога, объявляются бесконечными. Однако если откинуть веса ребер, то полученный обыкновенный граф может использоваться для кластер-анализа через отыскание компонент и клик – практически ничего хорошего из этого не получается, так как такой пороговый граф – слишком грубое приближение к структуре данных. Значительно лучше в этом плане работает подход на основе максимального покрывающего дерева, описанный выше в данном разделе. Но и он не обязательно приводит к хорошим результатам.

Дело в том, что этот метод может приводить к одному-двум очень большим кластерам объектов, тянущихся друг за другом (рис. 12), так что остальные объекты оказываются одиночками.

В этом плане могут представлять интерес графотеоретические конструкции, описанные в разделе 1.4.1 в связи с построением отдельных кластеров.

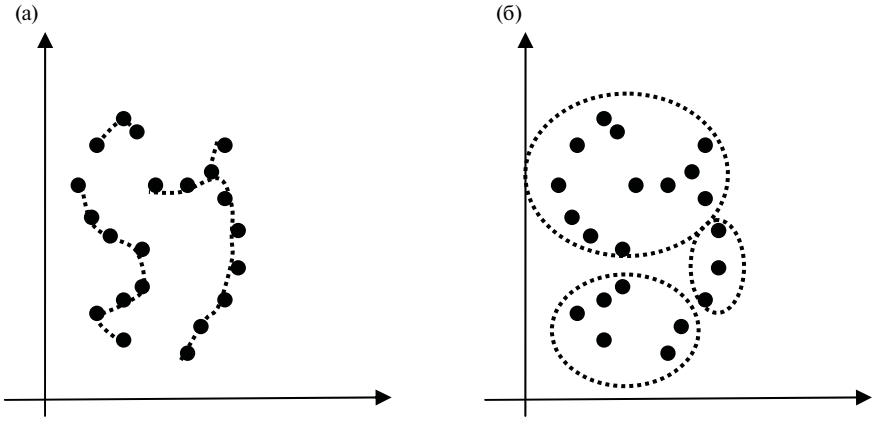

Рис. 12. Кластеры на одном и том же множестве:
по методу ближнего соседа (а) и методу *к*-средних (b)

Одно из наиболее интересных кластерно-графовых понятий – плотность подграфа, определяемая как отношение числа дуг в подграфе, полученном сужением основного множества вершин *I* до его подмножества *S*, и числа элементов в множестве *S*. Это та самая величина, которая максимизируется методом аппроксимационного кластера и является одним из слагаемых суммарного критерия (1.6). Подграф максимальной плотности можно построить за полиномиальное число шагов, что нетрудно перенести на случай взвешенного графа при неотрицательных связях. Это понятие иногда используется в биоинформатике.

1.4.3. Построение иерархий

1.4.3.1. Агломеративные методы

До относительно недавнего времени основные методы построения иерархий были агломеративными, т.е. такими, при которых вычисления начинались с тривиальных – одиночных – кластеров и шли итерациями, каждая из которых состояла в объединении двух наиближайших кластеров с последующим определением расстояния между вновь построенным кластером и остальными. Результат работы этого алгоритма зависит от формулы пересчета расстояний до вновь построенного кластера. Общая формула Ланса и Вильямса (1967) относится к случаю, когда расстояние до объединенного кластера

определяется исходя из расстояний до объединившихся частей. Хорошо известный пример – метод ближнего (дальнего) соседа, когда расстояние до объединенного кластера берется как минимум (максимум) расстояний до объединившихся кластеров. Эти два метода приводят к кластерам, которые в какой-то мере аналогичны графтеоретическим концепциям компоненты связности (ближний сосед) и клики (дальний сосед). Популярен также метод Уорда (1963), который использует разность значений критерия метода κ-средних (взвешенная дисперсия данных) на исходном и «объединенном» разбиениях. Оказывается, расстояние между кластерами по этому методу равно

$$dw(S_f, S_g) = \frac{N_f N_g}{N_f + N_g} d(c_f, c_g),$$

где d – квадрат Евклидова расстояния между центрами кластеров, а N_1 и N_2 – их численности. Именно коэффициент при расстоянии определяет популярность и, в какой-то мере, проблематичность критерия κ-средних: значение коэффициента определяется произведением $N_1 N_2$; оно минимально при максимальной разнице между N_1 и N_2, заставляя тем самым, при прочих равных, присоединять маленькие кластеры к большим (Mirkin 2005).

Агломеративные методы, к сожалению, используют все попарные расстояния для отыскания минимума, что может существенно осложнить вычисления на больших данных. Правда, имеется довольно широкий класс расстояний, для которых область поиска минимумов можно сильно ограничить [Murtagh 1983].

1.4.3.2. Дивизимные методы

В последнее время все большей популярностью пользуются дивизимные методы. Эти методы делают кластеры путем разбиения больших кластеров на меньшие части, начиная со всего множества. Удобство – процесс деления можно в любой момент остановить. Три популярных дивизимных метода – (1) бисекция κ-средних (раздвоение), (2) бисекция главной компоненты и (3) концептуальный кластер-анализ.

Наиболее часто применяемый из них – (1) бисекции κ-средних или раздвоение, который является адаптацией того же расстояния Уорда для дивизимной схемы, – этот последний факт, по-видимому,

ускользнул от внимания разработчиков, хотя и был доказан в [Mirkin 1996]. Согласно этому методу, критерием разбиения данного кластера S на части S_1, S_2 является максимизация расстояния Уорда, для чего вполне годится метод κ-средних при $K = 2$ ([Mirkin 1996, 2005], несмотря на то, что впервые он был описан в 1996 г., популярность метод получил только после публикации [Zhang et al. 2005], которые описывают его как оригинальный метод, и без ссылки на критерий). В последнее время также приобрел популярность метод PDDP, основанный на разделении данных, спроецированных на направление главной компоненты (2) [Boley 1998]. Этот метод сначала находит главную компоненту (первый сингулярный вектор матрицы данных), после чего данные разбиваются на две части, соответствующие ее положительной и отрицательной полуосям. Метод не зависит от инициализации, зато зависит – и сильно – от положения нуля, как и аномальные кластеры. В работе [Savaresi, Boley 2004] шаги этого метода и шаги метода бисекции κ-средних сравниваются, чтобы выяснить, почему эти два метода часто приводят к одинаковым решениям – по-моему, не очень убедительно. На самом деле объяснение этого явления вытекает из результатов, полученных в [Mirkin 1996]: оказывается, критерий Уорда – не что иное, как «тернарная» аппроксимация критерия метода главных компонент, используемого [Boley 1998, 2004]. Значит, эти два метода по-разному аппроксимируют одну и ту же проблему максимизации отношения Рэлея, рассматриваемого в методе главных компонент, чтобы найти бинарный вектор, определяющий раздвоение рассматриваемого кластера, а именно метод бисекции κ-средних аппроксимирует решение в самом критерии, а метод PDDP аппроксимирует найденное оптимальное решение. Тот факт, что они аппроксимируют одну и ту же проблему, хотя и по-разному, и предопределяет сходство результатов.

Что касается так называемого концептуального кластеринга – метод известен еще с 1960-х годов, хотя название ему дал покойный Рышард Михальски (1986), который считал древовидную структуру моделью отдельного понятия; в настоящее время это скорее модель таксономии или даже онтологии. В настоящее время метод используется только с бинарными разбиениями. Концептуальный кластеринг отличается от остальных методов тем, что разделение кластера в нем осуществляется не по многомерному критерию, а по одному из признаков. Если признак x количественный, то две части, на ко-

торые разбивается кластер, отвечают предикатам «x > *a*» и «x < = *a*» для некоторого значения *a*. Еще проще возможные разбиения, если признак – категоризованный. Тогда две части раздела отвечают предикатам «x = *a*» и «x ≠ *a*» для некоторого *a*. В процессе вычислений осуществляется полный перебор всех кластеров – кандидатов для разбиения, всех признаков и всех значений *a* – их на самом деле не очень много, линейная функция от числа признаков, и выбирается то разделение, для которого суммарная ассоциация с существующими признаками максимальна. Получаемое «концептуальное» дерево имеет очень простую интерпретацию и, кроме того, выступает в качестве инструмента отбора информативных признаков – тех, по которым действительно шло разделение. При этом может, правда, возникнуть ситуация, когда число делений слишком велико, так как не удается добиться требуемой степени однородности кластеров – тогда дерево слегка укорачивают (pruning – подрезка), отрезая слишком длинные ветви. Степень ассоциации может измеряться так называемым корреляционным отношением (хотя часто используются и другие, специально изобретенные меры) в случае количественных признаков, или коэффициентами, основанными на таблице сопряженности между искомым разбиением и признаком, в случае неметрических признаков. Причем в последнем случае некоторые известные коэффициенты ассоциации, популярные в построении решающих деревьев, такие как индекс Джини и коэффициент сопряженности Пирсона, оказываются эквивалентны специальным случаям критерия квадратичной ошибки в методе *к*-средних, при условии, что отдельные категории представлены как числовые 1/0 признаки (их часто называют фиктивными или дамми) и подходящим образом нормированы (см. [Mirkin 2011]).

1.4.4. Бикластерный анализ

При заданной матрице данных $Y = (y_{iv})$, где $i \in I$ и $v \in V$, I – множество строк, V – множество столбцов, бикластер – это пара множеств (S, T), где $S \subset I$ – подмножество строк, а $T \subset V$ – подмножество столбцов, такое, что подматрица $Y(S, T) = (y_{iv})$, где $i \in S$ и $v \in T$, имеет какую-либо особенность, обычно – одинаковые строчки или даже одинаковые значения, свидетельствующие об определенной связи между S и T, хотя иногда рассматривают и более сложные зависимости – например рост значения, пропорциональный номеру столбца.

1.4.4.А. Лингвистический подход

Для матрицы данных $Y = (y_{iv})$, где $i \in I$ и $v \in V$, биразбиение образуется двумя разбиениями, $S = \{S_1, S_2, ..., S_K\}$ на множестве строк I и $T = \{T_1, T_2, ..., T_L\}$ на множестве столбцов V, так что каждый блок (S_k, T_l) – это бикластер ($k = 1,..., K$ и $l = 1,..., L$). Такая структура часто рассматривается на бинарных данных или таблицах сопряженности (перекрестных (см., например, [Nadif, Govaert 2005]). Браверман и др. (1974) предложили несколько более гибкую кластерную структуру, подходящую для произвольных данных вида объект – признак, в которой признаки (элементы множества V) организованы в классы разбиения $T = \{T_1, T_2, ..., T_L\}$ на V, однако строки разбиваются не один, а L раз, для каждого из классов T независимо друг от друга. Такая блок-структура задается L-классным разбиением T на V и множеством L разбиений $S^l = \{S^l_1, S^l_2, ..., S^l_K\}$ на I, каждое разбиение S^l в полосе столбцов множества T_l ($l = 1,..., L$). Оригинальность этого подхода состоит в том, что для построения блок-структуры используются не непосредственно элементы матрицы данных Y, а первая главная компонента каждого подмножества признаков T_l. Такая формулировка приводит к значительно большей интерпретируемости получаемых главных компонент, поскольку здесь они не обязательно ортогональны друг другу, в отличие от решений, получаемых обычным методом главных компонент (см., например, [Браверман и др. 1974; Браверман, Мучник 1983]). Один из наиболее практичных методов Бравермана, Мучника для построения блок-структуры состоит из двух этапов. На первом этапе ищется разбиение T на множестве признаков V так, чтобы минимизировать внутригрупповые разности между первой главной компонентой множества T_l и каждым признаком, входящим в группу T_l. На втором этапе ищется разбиение множества объектов на K кластеров по единственному признаку – первой главной компоненте, найденной на первом шаге для каждой группы $l = 1,..., L$. Эта методика названа авторами лингвистическим анализом, поскольку по идее кластеры объектов описываются в алфавите кластеров признаков. Показателен пример применения метода лингвистического анализа к множеству 85 стран, охарактеризованных тремя десятками признаков социально-экономического развития. Эти признаки были разбиты на $L = 2$ кластера, один из которых включал душевой доход и связанные с ним признаки, а второй – степень капитализации экономики. Пересече-

ние двух 5-классных разбиений, полученных по построенным двум главным компонентам, оказалось соответствующим известному математическому выводу оптимальной динамики в двухсекторной модели экономики: максимальное суммарное потребление в заданный период времени достигается с помощью политики переключения инвестиций – в начале периода все инвестиции направляются в сектор промышленности, а в конце – в сектор потребления (Браверман и др. 1974). Авторы рассматривали также версию метода с критерием наименьших модулей, что привело к интересной модификации центроидного метода факторного анализа.

Несмотря на удачный выбор критериев и методов, в международной литературе лингвистический подход остался практически неизвестным. В последнее время развивается несколько более общий подход "subspace clustering" (см., например, [Kriegel 2009]).

Ростовцев распространил свой подход к анализу отдельных бикластеров на задачу формирования блок-структуры для матрицы связи. При этом он успешно использовал свой критерий максимального различия между результатами, найденными на реальных и случайных данных, для автоматизации выбора как L, так и K (см. [Ростовцев 1985], а также [Миркин 1985]).

1.4.4.Б Построение бикластеров

Хотя первые работы по бикластерному анализу были выполнены еще в 70-е годы прошлого столетия, а сам термин был введен автором в 90-е годы (см. [Mirkin 1996], где описаны аппроксимационные методы и некоторые приложения – очень успешным был предварительный бикластер-анализ матрицы связей между признаками с целью выявления совокупностей признаков для проведения относительных группировок [Миркин 1985]), бикластерный анализ как «движение» начался со статьи Cheng, Church (2000), в которой аппроксимационная модель была модифицирована применительно к проблематике анализа данных об экспрессии генов: искомый бикластер (V, W) должен удовлетворять уравнениям

$$a_{ij} = a + b_i + c_j + e_{ij}, \quad i \in V, j \in W,$$

где величины a, b, c, так же, как и кластер (V, W), ищутся из условия минимизации суммы квадратов величин невязок e_{ij}. Ясно, что опти-

мальные значения $a + b_i$ и $a + c_j$ равны средним по строке и, соответственно, столбцу внутри бикластера *(V,W)*. За истекшие годы опубликованы десятки алгоритмов бикластерного анализа (см., например, обзоры [Prelic et al. 2006; Bozdag et al. 2010]), но почему-то среди рассматриваемых методов нет наших, более ранних, методов, основанных на присоединении/исключении единственной строки, хотя они вполне конкурентоспособны, как показывает недавняя работа [Mirkin, Kramarenko 2011]. В этой последней работе задача бикластерного анализа обобщена на случай трех взаимосвязанных множеств и в этом виде представляет непосредственный интерес для анализа решений в ситуациях, когда для каждого объекта указано подмножество объектов, примерно эквивалентных по предпочтению, причем эти подмножества задаются эмпирически и не обязательно согласованы.

2. Принятие решений в кластер-анализе

В проблеме кластер-анализа нет никакого явно формализуемого критерия, а все формализации субъективны и не привязаны к модельным построениям. Поэтому под вопросом оказываются наиболее фундаментальные решения:

А. Выбор совокупности объектов.

Б. Выбор признаков и шкал их измерения.

В. Выбор меры близости между объектами.

Г. Выбор вида кластерной структуры и критерия ее отыскания.

Д. Выбор метода и его параметров.

Е. Выбор «грубости» кластерной структуры, например, числа кластеров.

Ж. Выбор из решений, полученных различными алгоритмами.

З. Валидация, т.е. проверка и установление области «доверия» результатам.

И. Интерпретация результатов.

К. Выводы из результатов и принятие решений на их основе.

К счастью, во многих случаях характер решений очевиден из структуры данных или условий задачи, а некоторые проблемы удается устранить, поместив проблематику в единообразный контекст. Так, на-

пример, для табличных данных перевод проблематики в аппроксимационное моделирование практически устраняет проблему выбора меры близости, переводя ее в легче анализируемую проблему выбора критерия аппроксимации, в которой иногда удается показать, что метод наименьших квадратов вполне может быть взят за основу.

Из упомянутых выше проблем мы остановимся на следующих:

– выбор числа кластеров (относится к Е);
– консенсус (Ж);
– наличие нескольких критериев (Г);
– валидация кластеров (З);
– интерпретация кластеров (И).

2.1. Число кластеров

На первый взгляд кажется, что проблема выбора числа кластеров не поддается никакому общему решению: см., например, рис. 13, на котором одни и те же объекты образуют четыре кластера слева, а справа два – только потому, что рост стали измерять в метрах, а не в футах.

Однако это не совсем так: при наличии многих объектов кластеры ассоциируются с местами, где объектов больше, чем в других местах, а это вполне возможно операционально установить – конечно, в пределах, определяемых естественными возможностями грануляции, т.е. уровня различимости, измерений.

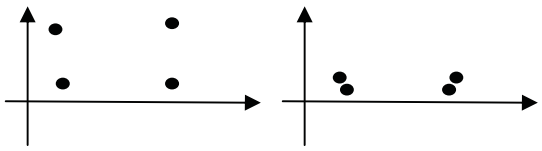

Рис. 13. Изменение кластерной структуры при укрупнении единицы измерения вертикальной шкалы

В настоящее время разумных предложений о том, как надо определять число кластеров, несколько десятков (см. их таксономию на рис. 14 из обзора [Mirkin 2011a], где также можно найти ссылки на источники).

Из них мы остановимся на трех:

– суммарная дисперсия по Хартигану;

– аномальные кластеры;
– остановка процесса дивизимного деления.

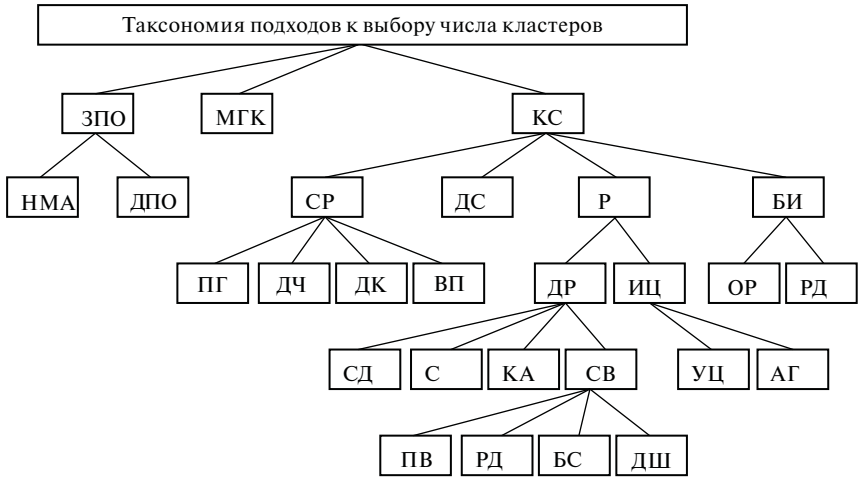

Рис. 14. Таксономия подходов к выбору числа кластеров (см. [Mirkin 2011a])

Здесь ЗПО – Знания о прикладной области; делится на НМА – Непосредственная модификация алгоритма и ДПО – Доработка после обработки. Другие две ветки: МГК – Моделирование генерации кластеров и КС – Кластерные структуры. Последняя разделяется на СР – Смеси распределений, Р – Разбиения, БИ – Бинарные иерархии и ДС – Другие структуры. СР, в свою очередь, подразделяется на ПГ – Проверку гипотез, ДЧ – Дополнительные члены в критерии максимального правдоподобия, ДК – Дополнительные критерии и ВП – Взвешенное правдоподобие. БИ разделяется на ОР – Остановка по критерию и РД, Разрез построенного дерева, а Р состоит из ДР – Доработка разбиения и ИЦ – Инициализация центроидов. ИЦ, в свою очередь, разделено на УЦ – Удаленные центроиды и АГ – Аномальные группы. ДР включает СД, подход, основанный на Суммарной дисперсии, С – Структурный подход, КА – Консенсус ансамбля разбиений, и СВ – подход, основанный на Случайных выборках. Этот последний состоит из ПВ – Подвыборка, РД – Разделение данных, БС – Бутстраппинг и ДШ – Добавление шума.

2.1.1. Суммарная дисперсия по Хартигану

Представим себе, что имеется K^* хорошо разделенных кластеров и проследим за поведением критерия суммарной дисперсии (1.8), он же (1.10), по методу k-средних. Очевидно, что при искомом числе кластеров $K < K^*$ любые локально оптимальные разбиения на K и $K + 1$ кластеров будут представлять собой это естественное разбиение,

в котором какие-то кластеры объединены в один. При этом значения критерия W_{K+1} и W_K на этих разбиениях будут сильно отличаться, даже если разбиение на K кластеров получено объединением каких-то из кластеров разбиения на $K + 1$ кластер. С другой стороны, если $K > K^*$, то оба K– и $(K + 1)$-кластерные разбиения будут совпадать с «естественным» разбиением на K^* кластеров, в котором какие-то кластеры случайно разбиты на части, так что W_K и W_{K+1} не будут сильно различаться.

На основе подобного рассуждения получается «среднепотолочный» критерий Hartigan (Hartigan 1975, p. 91):

$$H_K = (W_K/W_{K+1} - 1)(N-K-1), \qquad (2.1)$$

в котором N – число объектов, а W_K и W_{K+1} – минимальные значения критерия суммарной дисперсии (1.8) при разбиениях на K и $K + 1$ кластеров. Разумеется, в реальных приложениях эти значения неизвестны; в качестве оценок берутся минимальные значения критериев, полученные после 100 (или другого числа) применений метода κ-средних, начиная со случайно выбранных объектов в качестве центроидов. Согласно правилу Хартигана, надо последовательно разбивать множество на все большее число кластеров K, до тех пор, пока H_K не станет меньше 10 – именно это K и надо брать в качестве оценки K^*. Оказалось, что в экспериментах с круглыми и вытянутыми, перекрывающимися и нет сгенерированными кластерами, описанных в [Chiang, Mirkin 2010], это правило оказалось лучшим из совокупности девяти наиболее популярных критериев.

2.1.2. Аномальные кластеры

Метод аномальных кластеров был описан в разделе 1.4.2.2.B: каждый кластер начинается с самого дальнего от 0 – аномального – объекта в качестве центроида аномального кластера, после чего объекты итеративно распределяются между аномальным центроидом и 0. Полученный аномальный кластер изымается, и процесс начинается на оставшемся множестве – без изменения положения нуля. После того как все объекты распределены таким образом, центроиды кластеров, содержащих больше одного объекта, используются для инициализации метода κ-средних – при этом число кластеров определяется автоматически. В экспериментах (см. [Chiang, Mirkin 2010]) этот метод хорошо работал при относительно удаленных центроидах, хотя бы и

вытянутых, кластеров. При сближении центроидов число аномальных кластеров быстро начинает превышать число сгенерированных кластеров. По данным [Amorim, Mirkin 2011], а также некоторым другим сведениям, этот метод хорошо работает также при числе кластеров K, известном заранее так, что отбираются центроиды K первых самых многочисленных аномальных кластеров.

2.1.3. Функция плотности без минимумов

Недавно Tasoulis et al. (2010) предложили и экспериментально обосновали критерий остановки процесса разделения кластеров, когда плотность распределения в кластере вдоль его оси главной компоненты лишена минимумов (либо вогнута, либо монотонна). При этом в качестве оценки плотности бралась хорошо известная в теории распознавания образов сумма сферических Гауссианов, построенных в каждой наблюденной точке. Единственный параметр здесь – так называемое окно Парзена, стандартное отклонение Гауссиана берется равным $\sigma(4/N)^{1/5}$, где σ – стандартное отклонение данных на оси. В дипломной работе бакалавра Е.В. Ковалевой (2010) этот метод перенесен на еще более успешный подход проецирования кластеров не на главную ось, а на случайные направления, число которых, правда, должно быть порядка десяти.

2.2. Консенсус

Все острее осознаваемая проблема современного этапа развития кластер-анализа – слишком много доступных программ, реализующих разные методы, которые на одних и тех же данных формируют разные кластерные структуры. Самый типичный случай: метод κ-средних применен многократно к одним и тем же данным, при различном числе классов и исходя из различных начальных центроидов. Получаются сотни и тысячи кластерных решений. Какую структуру принять в качестве финального решения?

Кластерный консенсус – это агрегирование множества кластерных структур в единую кластерную структуру, наилучшим образом их представляющую. Эта тематика особенно популярна в анализе данных генной экспрессии (см., например, [Swift et al. 2004; Xiao, Pan 2007]). В последнее время становится популярным еще один термин: ансамбль кластеринг (см. обзор [Topchy et al. 2005]). Проблема кла-

стерного консенсуса для случая, когда кластерные структуры – разбиения впервые рассматривалась Б. Миркиным как в рамках аксиоматического, так и аппроксимационного подходов (см. [Mirkin 1974б]; аксиоматический анализ для более широкого класса структур описан в [Day, McMorris 2003]). Согласно аппроксимационному подходу, проблема ставится так: при заданных n разбиениях S_1, ..., S_n на I найти такое разбиение R, которое минимизирует $\Sigma_t d(R, S_t)$, где суммирование производится по всем t от 1 до n. Оказывается, проблема может быть эквивалентно переформулирована в терминах так называемой консенсус-матрицы M коэффициентов сходства между элементами множества I: обозначим через m_{ij} число таких разбиений S_t, в которых элементы $i, j \in I$ находятся в одном и том же классе и определим пороговое значение $\lambda = n/2$. Тогда оптимальное разбиение R максимизирует суммарное сходство внутри классов после вычитания порога λ (см. [Mirkin 1974a, 1974б])

$$f(R, \lambda) = \Sigma_k \Sigma_{ij \in Rk} (m_{ij} - \lambda) = \Sigma_k \Sigma_{ij \in Rk} m_{ij} - \lambda \Sigma_k |S_k|^2. \qquad (2.2)$$

Введение вычитаемого порога делает выгодным объединение в один класс таких объектов i и j, для которых $(m_{ij} - \lambda) > 0$ – тогда $f(R, \lambda)$ увеличится, и разделение их по разным классам, если $(m_{ij} - \lambda) < 0$ – тогда $f(R, \lambda)$ не уменьшится. Конечно, эти эффекты зависят от сходства между i, j и объектами, уже размещенными по классам. Критерий (2.2) иногда используется в анализе данных, хотя и без ссылок на первоисточник (см., например, [Swift et al. 2004]).

Существует следующий модельный тест (см. [Mirkin, Muchnik 2008]), для которого разрешающая способность критерия (2.2) недостаточна. Согласно этому тесту, имеется заданное разбиение $R = \{R_k\}$ из K классов $k = 1, 2, ..., K$, на множестве объектов I, например, по раскраске цветами радуги. Положим $n = K$ и определим S_k как разбиение на два кластера, R_k и $I - R_k$ ($k = 1, ..., K$), так что S_k соответствует бинарному атрибуту «есть цвет k или нет». Естественно ожидать, что кластерный консенсус для этих бинарных разбиений – не что иное как исходное разбиение R. Это действительно так при малых k: R максимизирует критерий (2.2), но только при $K \subset 4$. Если же $K > 4$, то оптимальным по критерию (2.2) будет универсальное разбиение U, состоящее из единственного кластера, состоящего из всех элементов I. Чтобы повысить разрешающую способность критерия, каждому разбиению S_t приписывается положительный вес, w_t, так

что кластерный консенсус R тоже должен иметь определенный вес, w, либо задаваемый экспертно, либо отыскиваемый согласно модифицированному критерию $\Sigma_t d(wR, w_t S_t)$, в котором расстояние d обобщено до суммы квадратов разностей между взвешенными матрицами отношений эквивалентности, соответствующих разбиениям S_t. Если в тестовом примере взять w равным $2(K-1)/K$ (при всех $w_t = 1$), так что значение $\lambda = (K-1)/K$, или же оптимальному значению, то, действительно, исходное разбиение R будет (взвешенным) кластерным консенсусом при любом K.

Этот аппроксимационный подход получил дальнейшее развитие в работе Миркина и Мучника (1981), в которой использовано более привычное представление разбиений в терминах матриц инциденций «объект – класс», а не в терминах квадратных матриц отношений эквивалентности. Для заданного разбиения $S = \{S_1, S_2, ..., S_K\}$ бинарная $N \times K$ матрица инциденций $X = (x_{ik})$ определяется условием: $x_{ik} = 1$, если $i \in S_k$ и $x_{ik} = 0$ – в противном случае. Эта матрица определяет линейное пространство $L(X)$, натянутое на столбцы – множество векторов Xa для всевозможных k-мерных a, причем оператор ортогонального проецирования на $L(X)$ определяется формулой $P_X = X(X^T X)^{-1} X$. Эта матрица P_X – квадратная размера $N \times N$, так что она может рассматриваться как матрица сходства на множестве I. Нетрудно видеть, элементы этой матрицы определяются условием $p_{ij} = 0$, если i и j принадлежат разным классам S, и $p_{ij} = 1/|S_k|$, если $i, j \in S_k$. Миркин и Мучник (1981) рассматривают два способа определения консенсус-разбиения. Для заданных n разбиений $S_1, ..., S_n$ на I, чьи матрицы инциденций обозначены $X_1, ..., X_n$, найти такое разбиение R с его матрицей инцидентности Z, которое минимизирует

$$\Sigma_t \|X_t - P_Z X_t\|^2 \text{ или } \Sigma_t \|Z - P_{X_t} Z\|^2$$

по всем возможным матрицам инциденций Z, где $\|D\|^2$ – обычная Евклидова норма, возведенная в квадрат, т.е. сумма квадратов всех элементов матрицы D.

Эти критерии приводит к так называемым порождающему и порождаемому консенсус-разбиениям соответственно. Действительно, согласно первому критерию, консенсус-разбиение сравнивается с остальными в его собственном пространстве $L(Z)$, а второй критерий переводит его в пространства каждого из исходных разбиений. Миркин и Мучник (1981) доказали, что эти критерии естественным

образом могут быть переформулированы в терминах сходства как между объектами, так и признаками. Оказывается, первый критерий эквивалентен стандартному в кластер-анализе критерию минимизации квадратичной ошибки в пространстве бинарных признаков, представленных столбцами матриц X_t, тогда как в терминах сходства между объектами ему эквивалентен другой популярный критерий – максимизации суммарных взвешенных связей внутри кластеров:

$$g(R) = \Sigma_k(\Sigma_{ij \in Rk} b_{ij})/|R_k|, \qquad (2.3)$$

где $R = \{R_1, R_2, ..., R_K\}$ – искомое консенсус-разбиение, а b_{ij} – сумма (i, j)-х элементов проекционных матриц на пространства $L(X_k)$, $k = 1, ..., K$.

Критерий (2.3) имеет несколько других интерпретаций.

Этот критерий, в случае произвольных матриц сходства, оказался лучшим из нескольких рассматриваемых интуитивно похожих критериев в экспериментах [Braverman et al. 1971].

Коэффициент сходства между объектами b_{ij}, определенный как сумма величин, обратных частотам тех категорий, которые совпадают на i и j, рассматривался еще в 1930-е годы энтомологом Е.С. Смирновым (1898–1972), заведующим кафедрой энтомологии Московского государственного университета с 1940 г., одного из первых энтузиастов так называемой нумерической таксономии, о которой ему удалось в конце концов опубликовать монографию Смирнов (1969).

Критерий порождаемого консенсус-разбиения имеет форму критерия (2.2), хотя и с другими коэффициентами сходства и порогом (см. [Миркин, Мучник 1981]).

Как отмечено выше, проблема кластерного консенсуса возникает в ситуациях двух разных типов – для согласования результатов разных кластерных процедур и для формирования «внутреннего» фактора по признакам, представленным разбиениями. Интуитивно критерий (2.2) больше подходит к первому случаю, а критерий (2.3) – ко второму, однако до сих пор не разработано теоретических представлений, позволяющих выразить это в явной форме, что остается делом будущего.

В последнее время эта проблематика привлекла внимание международного сообщества (см., например, [Strehl, Ghosh 2002; Swift et al. 2004; Topchy et al. 2005; Xiao, Pan 2007]), но кроме одного-двух эв-

ристических алгоритмов и вероятностной (а значит, трудно оцениваемой) модели Topchy et al. (2005), ничего интересного пока не предложено.

2.3. Валидация кластеров

Под валидацией кластеров понимают проверку их обоснованности. Различают два типа валидации: внутреннюю – по тому, насколько кластеры соответствуют данным, и внешнюю – по тому, насколько кластеры соответствуют информации, не учитывавшейся при их построении, но известной специалистам –такого рода информация обычно представляется в виде разбиения.

Казалось бы, для внутренней валидации можно использовать вклад кластеров в разброс данных $B(S, c)$ в формулах (1), (2) и (3) раздела 1.3.В или какой-либо производный критерий. Но нет, в литературе используются самые разнообразные индексы, выражающие традиционную идею, что внутри связь тесная, а между – нет, но не те, которые оптимизируются кластерными алгоритмами. Наиболее популярным является индекс Дэвиса – Болдина (Davies, Bouldin 1979), который можно определить следующим образом. Охарактеризуем относительный разброс в двух кластерах как полусумму средних расстояний их элементов до центров, деленную на расстояние между центрами. Охарактеризуем разброс кластера максимальной величиной его относительного разброса (относительно других кластеров). Тогда индекс Дэвиса – Болдина – не что иное, как средний разброс кластеров.

Используются также ширина силуэта (Kaufman, Rousseeuw 1990)) и индекс Данна (Dunn 1974), а также много других, менее популярных индексов (см., например, [Maulik, Bandyopadhyay 2002]). Для валидации нечетких кластеров применяется сумма квадратов степеней принадлежности объектов кластерам и различные производные критерии, поскольку эта величина характеризует степень близости объектов соответствующим центроидам в методе *c*-средних (см. раздел 1.4.2.2.Б) – одна из недавних работ с обзором на эту тему – Wu et al. (2009).

Что касается внешних критериев валидации, то наиболее популярным является так называемый отрегулированный индекс Рэнда – центрированная и нормализованная версия расстояния между разбиениями, аксиоматически введенного Миркиным и Черным (1970),

или, что то же самое – индекса Рэнда (1971), введенного в США, конечно, без ссылок. Этот коэффициент выражается формулой [Hubert, Arabie 1985]:

$$ARI = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\binom{N_{kl}}{2} - \left[\sum_{k=1}^{K}\binom{N_{k+}}{2}\sum_{l=1}^{L}\binom{N_{+l}}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_{k=1}^{K}\binom{N_{k+}}{2} + \sum_{l=1}^{L}\binom{N_{+l}}{2}\right] - \left[\sum_{k=1}^{K}\binom{N_{k+}}{2}\sum_{l=1}^{L}\binom{N_{+l}}{2}\right]/\binom{N}{2}}$$

где $\binom{N}{2} = \frac{N(N-1)}{2}$, а N с индексами выражает численности классов двух разбиений – кластерного и целевого, а также их пересечения. Чем выше значение этого индекса, тем больше соответствие между кластерами и целевым разбиением. Несмотря на то что формально индекс не зависит от чисел кластеров в разбиениях, реально он довольно сильно зависит от них. Mirkin (1996) пытается вычленить единые идеи в довольно большом количестве индексов, определенных для различных кластерных структур.

В целом вопрос о валидации остается в значительной степени открытым из-за отсутствия четко сформированных целей кластер-анализа.

2.4. Наличие нескольких критериев

Хотя кластер-анализ в силу неопределенности целей является плодотворной почвой для многокритериального подхода, эта ниша, начатая работой Ferligoj, Batagelj (1992), практически не развивалась до самого последнего времени. Работы Brusco, Stainley (2009) и Caballero et al. (2010) могут рассматриваться как представительный срез развития, которое ограничивается в основном чисто техническими аспектами. Отмечается, что можно оптимизировать один критерий по нескольким матрицам связи или несколько критериев – по одной матрице связи, после чего фиксируется три-четыре конкретных формулы и производится вычислительный эксперимент по построению и интуитивному анализу границы Парето. Что-либо другое здесь делать трудно, поскольку никаких соображений о сравнительном весе отдельных критериев нет и быть не может – все критерии и меры носят исключительно эвристический характер.

На наш взгляд, основной интерес в этом направлении может представлять формализация действительно различных критериев кластер-анализа, прежде всего внутренней и внешней валидации из предыдущего раздела, сбалансированности размеров кластеров, интерпретируемости кластеров и пр. – это, вероятно, дело будущего.

2.5. Интерпретация кластеров

В литературе мало внимания уделяется методам интерпретации. В соответствии с материалами [Mirkin 2011] и дальнейшими изысканиями автора, можно говорить о пяти уровнях интерпретации кластеров:

(a) кластерные центроиды относительно средних по всей совокупности;

(b) представители кластеров;

(c) вклады пар кластер – признак в разброс данных;

(d) концептуальное описание кластеров в терминах признаков, использованных при их построении;

(e) концептуальное описание кластеров в системах знаний.

Проиллюстрируем их на примере данных табл. 1 и 2 из раздела 1.1.1.

Таблица 3. Стандартизованные данные из табл. 2 для кластера компаний *A*. Ближайший к центроиду объект: Ant по расстоянию, Ave по скалярному произведению (оба выражены в промилле/тысячных).

	Доход	**Кап.**	**Пос.**	**ЭТ**	**Энер**	**Про**	**Торг**	**Расст**	**Ск. пр.**
Центр ст.	0.10	0.12	–0.11	–0.63	0.17	–0.02	–0.14		
Центр ре.	24.10	39.23	2.67	0.00	0.67	0.33	0.00		
Ave	–0.20	0.23	–0.33	–0.63	0.36	–0.22	–0.14	222	**524**
Ant	0.40	0.05	0.00	–0.63	0.36	–0.22	–0.14	**186**	521
Ast	0.08	0.09	0.00	–0.63	–0.22	0.36	–0.14	310	386

(a) Центроид.

В табл. 3 приводятся нормированные (делением на размах) и центрированные данные об *A*-кластере из этой таблицы. Первая строка характеризует центроид, как в реальных шкалах, так и стандартизо-

ванных. Последнее находится в верхней строке и довольно неплохо характеризует кластер – он на уровне среднего, отличаясь на 10–15% в ту или иную сторону по всем признакам, кроме ЭТ – здесь отклонение (от среднего!) 63%. Реальное значение равно 0 – ЭТ не применяется; и это значение, действительно, выделяет А-кластер из остальных.

(b) Представитель

Центроид состоит из средних арифметических и может не совпадать ни с каким объектом. Поэтому в качестве представителя обычно выбирается объект, ближайший к центроиду по расстоянию, используемому в критерии метода *к*-средних, обычно – Евклидовому. Однако формула (3) из раздела 1.3.В приводит к другой рекомендации – выбирать такой объект, на котором максимально скалярное произведение с центроидом. Табл. 3 как раз дает пример кластера, в котором эта рекомендация приводит к другому выбору, Ave, а не Ant, ближайшего по Евклидовскому расстоянию, по-видимому, из-за того, что знаки компонент первого вектора лучше соответствуют знакам компонент центроида. Вообще, вопрос о представителе смыкается с вопросом о типе. Для ряда эмпирических дисциплин, таких как минералогия, тип – это «усредненный» представитель. Для гуманитарных дисциплин, таких как литературоведение, тип – это, скорее, носитель идеальных характеристик, представляющих скорее экстремальные или даже несуществующие значения. Как показывает история, – вспомним, например, концепцию «социалистического реализма» – разные концепции типа могут оказаться иногда очень важными. Любопытно, что понятие идеального типа в определенной степени возможно моделировать в рамках нечеткого кластер-анализа (см. [Satarov, Mirkin 1990; Nascimento 2005]), что может оказаться полезным при значительных изменениях выборки – характеристики «усредненного» представителя плывут в соответствии с изменениями средних характеристик, тогда как идеальный тип остается стабильным [Nascimento 2005].

(c) Вклады пар кластер – признак

Эти вклады определяются как аддитивные элементы суммарного вклада в формуле (1) раздела 1.3.В, так что вклад пары (k, v), где k – кластер, а v – признак, равен $c_{kv}{}^2 N_k$, где c_{kv} – элемент центроида, а N_k – численность кластера. Интуитивный смысл квадрата при компоненте центроида иллюстрируется рис. 15: признак тем важнее, чем дальше его внутрикластерное среднее от общего среднего.

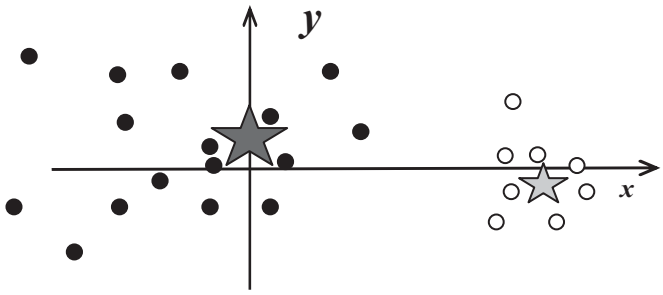

Рис. 15. Вклады признаков *x* и *y* в группу пустых кружков пропорциональны квадратам разностей между их центром (малая звезда) и общим центром (большая звезда), так что вклад *x* значительно выше, чем вклад *y*

Таблица 4. Центроиды и вклады пар признак – кластер для кластеров *A*, *B* и *C* данных о компаниях из табл. 2

Элемент		Доход	Кап.	Пос	ЭТ	Энер	Пром	Торг	Всего
Центроиды в стандартиз. шкалах	A	0.10	0.12	–0.11	–0.63	0.17	–0.02	–0.14	
	B	–0.21	–0.29	–0.22	0.38	–0.02	0.17	–0.14	
	C	0.18	0.24	0.50	0.38	–0.22	–0.22	0.43	
Вклады кластеров	A	0.0 3	0.05	0.04	1.17	0.09	0.00	0.06	1.43
	B	0.14	0.25	0.15	0.42	0.00	0.09	0.06	1.10
	C	0.06	0.12	0.50	0.28	0.09	0.09	0.38	1.53
Суммарные вклады	Объ.	0.23	0.41	0.69	1.88	0.18	0.18	0.50	4.06
	Вкл. %	31.1	59.4	77.5	100.0	28.6	28.6	100.0	68.3
Показатели относит. вклада, %	A	16.7	29.5	18.5	258.1	59.9	0.0	49.5	
	B	101.2	191.1	90.2	120.2	0.0	77.7	64.2	
	C	31.7	67.0	219.5	58.5	56.7	56.7	297.0	

Вклады пар «кластер – признак» для «продуктовых» кластеров табл. 1 представлены в табл. 4. Пары кластер – признак, дающие наибольший вклад, могут быть определены относительно строк (кластеры) или столбцов (признаки). Особенно наглядны расчеты по столбцам, ибо нам известны суммарные вклады отдельных признаков в разброс данных, так что можно оценить долю каждого признака, объясненную кластерами (по признакам ЭТ и Торг достигнута доля в

100%, что свидетельствует о полной согласованности признаков и кластеров). В последней строке таблицы даны относительные оценки вклада как по строкам, так и по столбцам – результаты деления относительного вклада признака в кластере на относительные вклады признаков в разброс данных. Признаки, дающие наибольший вклад, те, где кластеры действительно отличаются. *A*-кластер отличается признаком ЭК, в *B*-кластере Кап сравнительно низка, а *C*-кластер характеризуется сразу двумя признаками – Торг и Пос.

(d) Концептуальное описание кластеров признаками

По-видимому, наиболее удачное концептуальное описание кластера достигается как конъюнкция утверждений о значениях отдельных признаков типа «*x* находится между *a* и *b*», «y между *c* и *d*» и т.п. Близкое к этому описание дается структурой решающего дерева типа той, что отыскивается с помощью методов концептуального кластер-анализа, где совокупность ветвей дерева, отвечающих данному кластеру, может интерпретироваться как дизъюнкция конъюнкций, соответствующих каждой ветви (см. раздел 1.4.3.2). Часто используют популярные методы построения классификационных решающих деревьев (входят в популярные пакеты программ анализа данных, см., например, [Breiman et al. 1984; Green, Salkind 2003; Quinlan 1993]), чтобы описать таким образом уже построенное разбиение или отдельные кластеры. Можно пытаться дать концептуальное описание каждого отдельного кластера через конъюнкцию признаковых интервальных предикатов типа «*x* находится между *a* и *b*», особенно учитывая, что легко вычисляются вышеописанные показатели вклада каждого отдельного признака в объяснение каждого отдельного кластера. Автор пытался использовать этот подход как на исходных признаках, так и на их арифметических комбинациях (суммы, разности, отношения, произведения), см. [Mirkin 2005], но, несмотря на отдельные успехи, в целом этот подход пока не привел к продвижениям. Вероятно, значительно больше удастся сделать как в прикладном, так и теоретическом плане, для кластер-анализа сложных объектов, таких как последовательности или изображения. Для таких объектов признаки можно формировать как числовые или логические характеристики их фрагментов. Подобные попытки в литературе существуют, но исключительно для частных задач или для очень специальных типов данных, так что развитие этого подхода остается делом будущего.

(e) Концептуальное описание кластеров в онтологиях

Можно считать общепризнанным, что дальнейшая автоматизация поддержки систем знания – необходимых элементов систем принятия решений невозможна без использования вычислительных онтологий. Под онтологией понимают обычно множество понятий, относящихся к определенной части реального мира или знания, и отношений между ними. Вначале в качестве отношений рассматривались различные логические продукции (импликации) и фактологические утверждения. Однако в настоящее время, в связи с использованием онтологий для моделирования реальных систем, на первый план выходят таксономические отношения «быть частью», «состоять из» и т.п. Таксономия представима в виде корневого дерева все более общих (если идти от листьев) или все более детальных (если идти от корня) понятий. Таковы библиотечные классификации, классификация понятий информатики всемирной Ассоциации вычислительных машин (ABM, англ. – Association of Computing Machinery ACM) и множество таксономий, составляющих SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) – систему таксономий, разрабатываемых в настоящее время для нужд медицины (см., например, http://www.connectingforhealth.nhs.uk~/systemsandservices/data/ uktc/snomed). Использование онтологий для интерпретации кластеров понятий началось недавно, прежде всего для анализа того, каким понятиям более высокого уровня соответствует тот или иной кластер. Для этого используется либо условная вероятность понятие/кластер, либо ее отличие от вероятности самого понятия (см., например, [Mansingh et al. 2011; Marinica, Guillet 2009]). В работе [Mirkin et al. 2010] интерпретацию предлагается проводить путем наиболее экономного «подъема» кластера понятий на более высокие уровни таксономии за счет игнорирования «мелких» несоответствий между множеством и структурой дерева таксономии.

Заключение: основные направления развития

В данном обзоре кратко охарактеризованы современные методы кластер-анализа. При этом даже в тех разделах, которые напрямую как бы не относятся к методам – как, например, форматы данных

или кластерных структур, на самом деле главным подразумеваемым содержанием остаются методы. За последние пару десятков лет кластер-анализ превратился из одного из разделов анализа данных в отдельное направление, напрямую связанное и с системами поддержки знаний, и с системами принятия решений, и структурированием деятельности и рынков. В заключение хочется сказать несколько слов о том, какими видятся тенденции и перспективы дальнейшего развития. Представляется, что главные направления развития – это, прежде всего:

А. Дальнейшее формирование языка, на котором можно связывать разные задачи и подходы. Если в 60–80-е годы прошлого столетия доминировали полуэмпирические методы формирования кластеров, то где-то с 90-х годов стал доминировать модельный подход, прежде всего аппроксимационный, который позволил связать языки представления количественной и качественной информации. В настоящее время идет процесс интеграции трех основных способов представления данных: вероятностные распределения, таблицы объект – признак и матрицы связи, прежде всего через использование преобразования Лапласа и ядерных функций связи. Вероятно, следующим этапом будет дальнейшее развитие языка описания данных, так что такие характеристики, как структура данных и сложность структуры данных наполнятся операциональным смыслом. Одновременно должны осуществляться процессы формирования адекватных языков поддержки знаний, онтологии и таксономии, в которых кластерные конструкции и их интерпретации станут неотъемлемой частью.

Б. Дальнейшее развитие методов поддержки и обработки данных сложной структуры, прежде всего текстов, сигналов и изображений. Такие понятия, как скрытая марковская модель или суффиксное дерево, должны получить дальнейшее развитие как вычислительное, так и математическое, с тем, чтобы превратиться в конструкции, напрямую осуществляющие функции кодирования-декодирования данных для принятия решений. Временные ряды должны получить существенную математическую поддержку, чтобы выйти из «подполья». Это же относится и к социальным сетям Интернета. Ждут адекватной формализации динамические кластеры. На наш взгляд, перспективны с этой точки зрения подходы, переносящие на динамический случай аппроксимационные модели типа (1.11) в разделе 1.3.В или,

наоборот, дискретизирующие некоторые модели уравнений математической физики.

В. Разработка эффективных методов обработки больших массивов данных и распределенных массивов данных. Здесь, прежде всего, конечно, слово за вычислительной математикой – задачи перевода матричных и графовых вычислений в сублинейный режим за счет применения случайных выборок находятся на переднем краю усилий. Однако и собственно «кластерные» задачи – декомпозиция пространств, адекватно отображающих кластеры; комбинирование кластерных структур; распространение кластерных решений – несколько наиболее очевидных направлений, в которых работают специалисты.

Г. Интегрирование кластер-анализа и классификационных построений в единую математическую теорию классификационных структур, позволяющую анализировать и решать вопросы автоматического формирования адекватных признаковых пространств и концептуальных описаний. Представляется, что основой такой теории должны стать понятия «архетипа» и «классификации», находящиеся в двойственном отношении друг к другу [Мейен, Шрейдер 1976]. Вероятно, это направление получит серьезный импульс тогда, когда начнется систематическая разработка проблемы кластер-анализа объектов сложной структуры и, соответственно, проблема автоматического порождения признаков выйдет на первый план.

Д. Комбинирование аппарата кластер-анализа и теории принятия решений в подсистему поддержки принятия решений, позволяющую автоматизировать выбор альтернатив и критериев их выбора.

Литература

Авен П.О., Ослон А.А., Мучник И.Б. (1988) Функциональное шкалирование. М.: Наука, 1988.

Айвазян С.А., Бухштабер В.М., Енюков И.С. и др. (1989) Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989.

Айвазян С.А., Бежаева З.И., Староверов О.В. (1974) Классификация многомерных наблюдений. М.: Статистика,1974.

Айзерман М.А., Браверман Э.М., Розоноэр Л.И. (1970) Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

Браверман Э.М., Дорофеюк А.А., Лумельский В.И. и др. Диагонализация матриц сходства и отыскание скрытых факторов // Вопросы расширения возможностей автоматов. 1971. № 1. С. 42–79.

Браверман Э.М., Киселева Н.Е., Мучник И.Б. и др. Лингвистический подход к задаче обработки больших массивов информации // Автоматика и телемеханика 1974. № 11. С. 73–88.

Браверман Э.М., Мучник И.Б. (1983) Структурные методы обработки эмпирических данных. М.: Наука, 1983.

Елкина В.Н., Загоруйко Н.Г. (1966) Об одном алфавите распознавания // Вычислительные системы. 1966. № 12.

Загоруйко Н.Г. (1999) Прикладные методы анализа данных и знаний. Новосибирск, 1999.

Заславская Т.И., Мучник И.Б. (1980) Социально-демографическое развитие села: региональный анализ. М.: Статистика, 1980.

Ковалева Е.В. (2010) Методы остановки в дивизимных алгоритмах кластер-анализа. ВКР НИУ ВШЭ. М., 2010.

Кузнецов Е.Н., Мучник И.Б. (1981) Монотонные системы для анализа организационных структур // Методы анализа многомерной экономической информации. Новосибирск: Наука, 1981. С. 71–83.

Куперштох В.Л., Миркин Б.Г., Трофимов В.А. (1976) Сумма внутренних связей как критерий качества классификации // Автоматика и телемеханика. № 3. С. 91–100.

Лбов Г.С., Пестунова Т.М. (1985) Группировка объектов в пространстве разнотипных переменных // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 141–149.

Левенштейн В.И. (1965) Двоичные коды с исправлением выпадений, вставок и замещений символов. 1965. № 163(4). С. 845–848.

Малишевский А.В. (1998) Качественные модели в теории сложных систем. М.: Физматгиз, 1998.

Мандель И.Д. (1988) Кластерный анализ. М.: Финансы и статистика, 1988.

Мгеладзе А., Гоциридзе Г. (2009) Кластер-анализ в исследовании организационных систем. Тбилиси, 2009.

Мейен С.В., Шрейдер Ю.А. (1976) Методологтческие аспекты теории классификации // Вопросы философии. 1976. № 12. С. 67–79.

Миркин Б.Г. (1985) Группировки в социально-экономических исследованиях. М.: Финансы и статистика, 1985.

Миркин Б., Черный Л. (1970) Об измерении расстояния между различными разбиениями конечного множества // Автоматика и телемеханика. № 5. С. 91–98.

Миркин Б., Мучник И. (1981) Геометрическая интерпретация критериев кластер-анализа // Методы анализа многомерных экономических данных. Новосибирск: Наука, 1981. С. 3–11.

Миркин Б., Ростовцев П. (1978) Метод формирования взаимосвязанных групп признаков // Модели агрегирования социально-экономических данных. Новосибирск, 1978. С. 107–112.

Мучник И.Б. (1974) Анализ структуры экспериментальных графов // Автоматика и телемеханика. № 9. С. 62–80.

Мучник И.Б., Ослон А.А. (1980) Построение фактора, аппроксимирующего матрицу связей // Автоматика и телемеханика. 1980. № 4. С. 89–96.

Орлов А.И. (1985) Общий взгляд на статистику объектов нечисловой природы // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 58–91.

Портал о корпоративных порталах: консалтинг, создание, внедрение и поддержка. URL: http://corportal.ru/ (дата обращения: 04.02.2011).

Ростовцев П.С. (1985) Алгоритмы анализа структуры прямоугольных матриц «пятна» и «полосы» // Анализ нечисловой информации в социологических исследованиях. М.: Наука. С. 203–212.

Сатаров Г.А., Миркин Б.Г. (1991) Метод аддитивных нечетких типов // Автоматика и телемеханика. № 5. С. 134–140 (Ч. 1.); № 7. С. 121–126 (Ч. 2).

Смирнов Е.С. (1969) Таксономический анализ. М.: МГУ, 1969.

Черняк Е.Л. (2010) Анализ текстовой информации для отображения на онтологию (для структурирования математических дисциплин, преподаваемых в ВШЭ). ВКР НИУ ВШЭ. М., 2010.

Amorim R., Mirkin B. (2011) Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering (Submitted, in the process of revision).

Apresian Y.D. (1966) An algorithm for finding clusters by a distance matrix, Computer. Translation and Applied Linguistics, 9: 72–79 (in Russian).

Arabie P., Boorman S.A., Levitt P.R. (1978) Constructing block models: how and why // Journal of Mathematical Psychology, 17: 21–63.

Bandyopadhyay S., Maulik U. (2002) An evolutionary technique based on K-means algorithm for optimal clustering in R^N // Information Sciences, 146: 221–237.

Bandelt H.-J., Dress A.W.M. (1989) Weak hierarchies associated with similarity measures – an additive clustering technique // Bulletin of Mathematical Biology, 51: 133–166.

Ben-Dor A., Shamir R., Yakhini Z. (1999) Clustering gene expression patterns // Journal of Computational Biology, 6, 3/4: 281–298.

Bie T., Cristianini N. (2006) Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems // Journal of Machine Learning Research, 7(Jul.): 1409–1436.

Biernacki C., Celeux G., Govaert G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood // IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7): 719–725.

Bock H.H. (1996) Probability models and hypotheses testing in partitioning cluster analysis // Clustering and Classification / P. Arabie, L. Hubert, G. De Soete (eds.). World Scientific Publ., River Edge (NJ), 377–453.

Bouguila N., Ziou D. (2007) High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on Minimum Message Length, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (10): 1716–1731.

Bozdogan H. (1994) Mixture-model cluster analysis and choosing the number of clusters using a new informational complexity ICOMP, AIC, and MDL model-selection criteria // Multivariate Statistical Modeling / H. Bozdogan, S. Sclove, A. Gupta et al. (eds.). Dordrecht: Kluwer, 1994. Vol. II. 69–113.

Bivand R., Pebesma E., Gómez-Rubio V. (2008) Applied Spatial Data Analysis with R, Springer.

Boley D. (1998) Principal direction divisive partitioning // Data Mining and Knowledge Discovery, 4: 325–344.

Bozdag D., Kumar A., Catalyurek U. (2010) Comparative analysis of biclustering algorithms, Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. N.Y.

Breiman L., Friedman J.H., Olshen R.A. et al. (1984) Classification and Regression Trees, Belmont, Ca: Wadswarth.

Brusco M.J., Steinley D. (2009) Cross validation issues in multiobjectiveclustering // British Journal of Mathematical and Statistical Psychology, 62: 349–368.

Caballero R., Laguna M., Martí R. et al. (2010) Scatter tabu search for multiobjective clustering problems // Journal of the Operational Research Society. 2010. P. 1–13

Cheng Y., Church G.M. (2000) Biclustering of expression data, Proceedings of 8th International Conference Intelligent Systems for Molecular Biology, 93–103.

Coppi R., D'Urso P., Giordani P. (2010) A fuzzy clustering model for multivariate spatial time series // Journal of Classification, 27 (1): 54–88.

Davies D.L., Bouldin D.W. (1979) A cluster separation measure // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1: 224–227.

Day W.H.E., McMorris F.R. (2003) Axiomatic Consensus Theory in GroupChoice and Bioinformatics. SIAM, Philadelphia, 2003.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological), 39 (1): 1–38.

Depril D., Van Mechelen I., Mirkin B. (2008) Algorithms for additive clustering of rectangular data tables // Computational Statistics and Data Analysis, 52: 4923–4938.

Dhillon I.S., Modha D.M. (2001) Concept Decompositions for Large Sparse Text Data using Clustering, Machine Learning, 42:1, 143–175.

Diday E. Orders and overlapping clusters by pyramids (1986) // Multidimensional Data Analysis / J. de Leeuw, W. Heiser, J. Meulman, F. Critchley (eds.). Leiden, DSWO Press, 201–234.

DiMaggio P., McAllister S.R., Floudas C.A. et al. (2008) Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies, BMC Bioinformatics, 9: 458.

Dunn J. (1974) Well separated clusters and optimal fuzzy partitions // Journal of Cybernetics, 4, 95–104.

Everitt B.S., Landau S., Leese M. et al. (2011) Cluster Analysis. 5th ed. Wiley, 2011.

Feldman R., Sanger J. (2007) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press.

Ferligoj A., Batagelj V. (1992) Direct multicriteria clustering algorithms // Journal of Classification, 1, 43–61.

Fiedler M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory // Czech. Math Journal. Vol. 25. P. 619–637.

Fisher D.W. (1987) Knowledge acquisition via incremental conceptual clustering // Machine Learning, 2, 139–172.

Florek K., Lukaszewicz J., Perkal H. et al. (1951) Sur la liason et la division des points d'un ensemble fini, Colloquium Mathematicum, 2, 282–285.

Gama J. (2010) Knowledge Discovery from Data Streams, Boka Raton, Florida, Chapman and Hall/CRC.

García-Escudero L., Gordaliza A., Matrán C. et al. (2010) A review of robust clustering methods, Advances in Data Analysis and Classification, 4, 2–3: 89–109.

Gower J.C., Ross G.J.S. (1967) Minimum spanning tree and single linkagecluster analysis // Applied Statistics, 18, 54–64.

Green S.B., Salkind N.J. (2003) Using SPSS for the Windows and Macintosh: Analyzing and Understanding Data, Prentice Hall.

Guo J., Hartung S., Komusiewicz C. et al. (2010) Exact algorithms and experiments for hierarchical tree clustering, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10).

Gusfield D. (1997) Algorithms on String, Trees, and Sequences, Cambridge University Press, 1997.

Hartigan J.A. (1972) Direct clustering of a data matrix // Journal of American Statistical Association, 67, 123–129.

Hartigan J.A. (1975) Clustering Algorithms. N.Y.: J.Wiley & Sons.

Holzinger K.J., Harman H.H. (1941) Factor Analysis. Chicago: University of Chicago Press, 1941.

Hubert L., Arabie P. (1985) Comparing partitions // Journal of Classification, 2, 193–218.

Huang Z., Ng M.K., Rong H. et al. (2005) Automated variable weighting in k-means type clustering // IEEE Transactions on Pattern Analysis and Machine Learning, 27(5), 657–668.

Jain A.K. (2010) Data clustering: 50 years beyond k-means // Pattern Recognition Letters, 651–666.

Jain A.K., Dubes R.C. (1988) Algorithms for Clustering Data, Englewood Cliffs, NJ, Prentice Hall.

Jenssen R., Erdogmus D., Principe J.C. et al. (2005) The Laplacian PDF Distance: A Cost Function for Clustering in a Kernel Feature Space // Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, 625–632.

Johnson D.S., Trick M.A. (eds.) (1996) Cliques, Coloring, and Satisfiability // DIMACS Series in Discrete mathematics and theoretical computer science. Vol. 26.

Johnson S.C. (1967) Hierarchical clustering schemes // Psychometrika, 32, 241–245.

Karplus K., Barrett C., Hughey R. (1998) Hidden Markov models for detecting remote protein homologies // Bioinformatics. 14 (10): 846–856.

Kaufman L., Rousseeuw P. (1990) Finding Groups in Data: An introduction to cluster analysis. Wiley, 1990.

Kettenring J. (2006) The practice of cluster analysis // Journal of Classification, 23, 3–30.

Kiselev M.V., Ananyan S.M., Arseniev S.B. (1999) LA – A Clustering algorithm with an automated selection of attributes, which is invariant to functional transformations of coordinates // Principles of Data Mining and Knowledge Discovery / J.M. Zytkow, J. Rauch (eds.). Third European Conference, PKDD 1999, Prague, 366–371.

Kriegel H.-P., Kroeger P., Zimek A. (2009) Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Transactions on Knowledge Discovery from Data (TKDD), 3, 1–58.

Lance G.N., Williams W.T. (1967) A general theory of classificatory sorting strategies: 1. Hierarchical Systems // Comp. Journal, 9, 373–380.

Lebart L., Morineau A., Piron M. (1995) Statistique Exploratoire Multidimensionnelle, Paris, Dunod.

Lo C., Yeung A. (2006) Concepts and Techniques of Geographic Information Systems, New Deli, Prentic Hall.

Lu Y., Lu S., Fotouhi F., Deng Y. et al. (2004) Incremental genetic algorithm and its application in gene expression data analysis, BMC Bioinformatics, 5, 172.

Luxburg U. von (2007) A Tutorial on Spectral Clustering. Statistics and Computing 17 (4): 395–416.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematics, Statistics and Probability. University of California Press, 281–297.

Manning C.D., Schiitze H. (1999) Foundations of Statistical Natural Language Processing Cambridge, MA: The MIT Press.

Mansingh G., Osei-Bryson K.-M., Reichgelt H. (2011) Using ontologies to facilitate post-processing of association rules by domain experts // Information Sciences 181, 419–434.

Marinica C., Guillet F. (2009) Improving post-mining of association rules with ontologies, The XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA).

Maulik U., Bandyopadhyay S. (2002) Performance evaluation of some clustering algorithms and validity indices // IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 1650–1654.

Milligan G.W. (1981) A Monte Carlo study of thirty internal criterion measures for cluster analysis // Psychometrika, 46, 187–199.

Ming-Tso Chiang M., Mirkin B. (2010) Intelligent choice of the number of clusters in K-Means clustering: an experimental study with different cluster spreads // Journal of Classification, 27(1), 3–40.

Mirkin B. (1987) Additive clustering and qualitative factor analysis methods for similarity matrices // Journal of Classification, 4, 7–31; Erratum, 6 (1989), 271–272.

Mirkin B. (1990) A sequential fitting procedure for linear data analysis models // Journal of Classification, 7, 167–195.

Mirkin B. (1996) Mathematical Classification and Clustering, Dordrecht-Boston-London, Kluwer Academic Publishers.

Mirkin B. (2005) Clustering for Data Mining: A Data Recovery Approach, Boca Raton Fl., Chapman and Hall.

Mirkin B. (2011) Core Concepts in Data Analysis: Correlation, Summarization, Visualization, London, Springer.

Mirkin B. (2011a) Choosing the number of clusters, WIRE Data Mining and Knowledge Discovery, to appear.

Mirkin B., Camargo R., Fenner T. et al. (2010) Similarity clustering of proteins using substantive knowledge and reconstruction of evolutionary gene histories in herpesvirus // Theoretical Chemistry Accounts: Theory, Computation, and Modeling 125, No. 3–6, 569–582.

Mirkin B., Kramarenko A. (2011) Approximation biclusters and triclusters for binary data // Proceedings of Thirteenth International Confe-

rence on Rough Sets, Fuzzy Sets and Granular Computing (RSFDGrC-11)", to appear.

Mirkin B., Muchnik I. (1996) Clustering and multidimensional scaling in Russia (1960–1990): A review // Clustering and Classification / P. Arabie, L. Hubert, G. De Soete (eds.). River Edge, NJ: World Scientific Publishing, 295–339.

Mirkin B., Muchnik I. (2002) Induced layered clusters, hereditary mappings, and convex geometries // Applied Mathematics Letters, 15, 293–298.

Mirkin B., Muchnik I. (2008) Some topics of current interest in clustering: Russian approaches 1960–1985 // Electronic Journal for History of Probability and Statistics, 4.

Mirkin B., Nascimento S., Fenner T. et al. (2010) Building fuzzy thematic clusters and mapping them to higher ranks in a taxonomy // International Journal of Software and Informatics, 4, 257–275.

Mitsa T. (2010) Temporal Data Mining, Chapman & Hall/CRC.

Muchnik I.B., Schwarzer L.V. (1990) Maximization of generalized characteristics of functions of monotone systems, Automation and Remote Control, 53, 1562–1572.

Murtagh F. (1983) A survey of recent advances in hierarchical clustering algorithms // The Computer Journal, 26, 354–359.

Nascimento S. (2005) Fuzzy Clustering via Proportional Membership Model, ISO Press.

Newman M.E.J. (2006) Modularity and community structure in networks, PNAS, 103 (23), 8577–8582.

Newman M., Girvan M. (2004) Finding and evaluating community structure in networks // Physical Review E, 69, 026113.

Ng A., Jordan M., Weiss Y. (2002). On spectral clustering: analysis and an algorithm // Advances in Neural Information Processing Systems / T. Dietterich, S. Becker, Z. Ghahramani (eds.).. MIT Press.

Pampapathi R. (2008) Annotated Suffix Trees for Text Modelling and Classification, PhD Birkbeck University of London, UK.

Pampapathi R., Mirkin B., Levene M. (2006) A suffix tree approach to anti-spam email filtering // Machine Learning, 65, 309–338.

Paterlini S., Krink T. (2006) Differential evolution and PSO in partitional clustering, Computational Statistics and Data Analysis, 50, 1220–1247.

Power D.J. (2007) A brief history of decision support sytems, DSSResources.com. URL: www.groupdecisionroom.nl (дата обращения: 30.01.2011).

Quinlan J.R. (1993) C4.5: Programs for Machine Learning, San Mateo: Morgan Kaufmann.

Prelic A., Bleuler S., Zimmermann P. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (9): 1122–1129.

Rand W.M. (1971) Objective criteria for the evaluation of clustering methods // Journal of the American Statistical Association, 66, 846–850.

Savaresi S.M., Boley D. L. (2004) A comparative analysis on the bisecting K-means and the PDDP clustering algorithms, 2004, Intelligent Data Analysis, IOS Press, 8 (4), 345–362.

Shepard R.N., Arabie P. (1979) Additive clustering: representation of similarities as combinations of overlapping properties // Psychological Review, 86, 87–123

Shi J., Malik J. (2000) Normalized cuts and image segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (8), 888–905.

Stanforth R., Mirkin B., Kolossov E. (2007) A measure of domain of applicability for QSAR modelling based on Intelligent K-Means clustering // QSAR & Combinatorial Science, 26 (7), 837–844.

Steinley D., Brusco M. (2007) Initializing K-Means batch clustering: A critical evaluation of several techniques // Journal of Classification, 24, 99–121.

Stepp R., Michalski R.S. (1986) Conceptual clustering of structured objects: A goal-orientied approach // AI Journal.

Strehl A., Ghosh J. (2002) Clustering ensembles – a knowledge reuse framework for combining multiple partitions // Journal of Machine Learning Research, 3, 583–617.

Swift S., Tucker A., Vinciotti V. et al. (2004) Consensus clustering and functional interpretaton of gene-expression data, Genome Biology, 5: R94.

Topchy A., Jain A.K., Punch W. (2005) Clustering ensembles: Models of consensus and weak partition // IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1866–1881.

Ward J.H. (1963) Hierarchical grouping to optimize an objective function // Journal of American Statist. Assoc., 58, 236–244.

Wu K.-L., Yang M.-S., Hsieh J.-N. (2009) Robust cluster validity indexes, Pattern Recognition, 42, 2541–2550.

Xiao G., Pan W. (2007) Consensus clustering of gene expression data and its application to gene function prediction // Journal of Computational & Graphical Statistics, 2007, 16 (3).

Xu R., Wunsch II D.C. (2009) Clustering, Wiley and Sons.

Zhao Y., G. Karypis G., Fayyad U. (2005) Hierarchical clustering algorithms for document datasets, Data Mining and Knowledge Discovery, 10 (2), 141–168.

Препринт WP7/2011/03
Серия WP7
Математические методы анализа решений
в экономике, бизнесе и политике

Миркин Борис Григорьевич

Методы кластер-анализа для поддержки принятия решений: обзор

Зав. редакцией оперативного выпуска *А.В. Заиченко*
Корректор *Е.Л. Качалова*
Технический редактор *Ю.Н. Петрина*

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета
Формат 60×84 $^{1}/_{16}$. Тираж 150 экз. Уч.-изд. л. 5,2
Усл. печ. л. 5,1. Заказ № . Изд. № 1341

Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»
Тел.: (499) 611-24-15

Для заметок

Для заметок