Экспериментальная психология, 2010, том 3, № 2, с. 67–85

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ОБРАБОТКИ ДАННЫХ ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

САВЧЕНКО Т. Н., *Институт психологии РАН, Москва*

В данной статье приводятся общие соображения о природе и целях кластер-анализа, а также сравнительный анализ и описание наиболее используемых методов кластерного анализа. Представлены стандартные методы, реализованные в наиболее часто используемых статистических пакетах, их развитие и усовершенствование, описаны оригинальные методы, отсутствующие в статистических пакетах.

Ключевые слова: многомерный анализ данных, классификация, кластер, модели латентных групп, агломеративный иерархический метод, метрика, дендрограмма, стратегия объединения, матрица расстояний, итерационная процедура, дендрит, оценка связанности, ассоциативный и цепной кластеры, кластер-коллекция, метод латентных классов.

Классификация является одним из фундаментальных процессов в науке. Прежде чем мы сможем понять определенный круг явлений и разработать принципы, их объясняющие, часто необходимо их предварительно упорядочить. Таким образом классификацию можно считать интеллектуальной деятельностью высокого уровня, которая необходима нам для понимания природы. Классификация – это упорядочение объектов по схожести. А само понятие схожести является неоднозначным. Принципы классификации также могут быть различными. Поэтому часто процедуры, используемые в кластерном анализе для формирования классов, основываются на фундаментальных процессах классификации, присущих людям и, возможно, другим живым существам (Классификация и кластер, 1980). Достаточно часто в психологии возникает необходимость проведения классификации множества объектов по множеству переменных. Для проведения такой многомерной классификации используются методы кластерного анализа. Группы близких по какому-либо критерию объектов обычно называются кластерами. Кластеризацию можно считать процедурой, которая, начиная работать с тем или иным типом данных, преобразует их в данные о кластерах. Многие методы кластерного анализа отличаются от других методов многомерного анализа отсутствием обучающих выборок, т.е. априорной информации о распределении соответствующих переменных генеральной совокупности. Методов кластерного анализа достаточно много, и далее будет описана их классификация.

Наибольшее распространение в психологии получили иерархические агломеративные методы и итерационные методы группировки. При использовании методов кластерного анализа достаточно сложно дать однозначные рекомендации по предпочтению применения тех или иных методов. Необходимо понимать, что получаемые результаты классификации не являются единственными. Предпочтительность выбранного метода и полученных результатов следует обосновать.

Кластерный анализ (КА) строит систему классификации исследуемых объектов и переменных в виде дерева (дендрограммы) или осуществляет разбиение объектов на заданное число удаленных друг от друга классов.

Методы кластерного анализа можно расклассифицировать на:

- внутренние (признаки классификации равнозначны);
- внешние (существует один главный признак, остальные определяют его).

Внутренние методы в свою очередь можно разделить на:

- иерархические (процедура классификация имеет древовидную структуру);
- неиерархические.

Далее, иерархические подразделяются на:

- агломеративные (объединяющие);
- дивизивные (разъединяющие).

Необходимость в использовании методов кластерного анализа возникает в том случае, когда задано множество характеристик, по которым тестируется множество испытуемых; задача состоит в выделении классов (групп) испытуемых, близких по всему множеству характеристик (профилю). На первом этапе матрица смешения (оценки людей по различным характеристикам) преобразуется в матрицу расстояний. Для подсчета матрицы расстояния осуществляется подбор метрики, или метода вычисления расстояния между объектами в многомерном пространстве. Если объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве. Возможность измерения расстояний между объектами в k-мерном пространстве вводится через понятие метрики.

Пусть объекты i и j принадлежат множеству M и каждый объект описывается k признаками, тогда будем говорить, что на множестве M задана метрика, если для любой пары объектов, принадлежащих множеству M, определено неотрицательное число d_{ij}, удовлетворяющее следующим условиям (аксиомам метрики):

1. Аксиома тождества: $d_{ij} = 0 \Leftrightarrow i \equiv j$.
2. Аксиома симметричности: $d_{ij} = d_{ji} \; \forall \; i, j$.
3. Неравенство треугольника: $\forall \; i, j, z \in M,$ выполняется неравенство $d_{iz} \le d_{ij} + d_{zj}$.

Пространство, на котором введена метрика, называется метрическим. Наиболее используемыми являются следующие метрики:

1. Метрика Евклида:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \; .$$

Эта метрика является наиболее используемой и отражает среднее различие между объектами.

2. Метрика нормированного Евклида. Нормализованные евклидовы расстояния более подходят для переменных, измеряемых в различных единицах или значительно различающихся по величине.

Если дисперсии по характеристикам отличаются друг от друга, то:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{S_k^2}} \; .$$

Если масштаб данных различен, например, одна переменная измерена в стэнах, а другая в баллах, то для обеспечения одинакового влияния всех характеристик на близость объектов используется следующая формула подсчета расстояния:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})^2}{x_{k\max}^2}}.$$

3. Метрика city-block (манхэттенская метрика, получившая свое название в честь района Манхэттен, который образуют улицы, расположенные в виде пересечения параллельных прямых под прямым углом; как правило, применяется для номинальных или качественных переменных):

$$d_{ij} = \sqrt{\sum_{k=1}^{n} | x_{ik} - x_{jk} |}.$$

4. Метрика на основе корреляции: $d_{ij} = 1 - |r_{ij}|$.

5. Метрика Брея-Картиса, которая также используется для номинативных и ранговых шкал, обычно данные предварительно стандартизируются:

$$d_{ij} = \sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{|x_{ik} + x_{jk}|}.$$

Расстояния, вычисляемые на основе коэффициента корреляции, отражают согласованность колебаний оценок, в отличие от метрики Евклида, которая определяет схожесть в среднем. Выбор метрики определяется задачей исследования и типом данных. Помимо приведенных выше методов, разработаны метрики для ранговых и дихотомических переменных и т.д. (во всех выше приведенных формулах i,j – номера столбцов; k – номер строки; d_{ij} – элемент матрицы расстояний; x_{ik}, x_{jk} – элементы исходной матрицы; n – количество объектов).

Наиболее используемый в психологии метод кластерного анализа – это *иерархический агломеративный метод,* который позволяет строить дерево классификации n объектов посредством иерархического объединения их в группы, или кластеры, все более высокой общности на основе заданного критерия, например, минимума расстояния в пространстве m переменных, описывающих объекты. В результате производится разбиение некоторого множества объектов на естественное число кластеров. Первоначально каждый элемент является классом, далее на каждом шаге происходит объединение ближайших объектов, и в результате все объекты образуют один класс.

Алгоритм агломеративного метода можно представить в следующем виде: на входе имеется матрица смешения, из которой строится матрица расстояния, либо матрица расстояния, полученная непосредственно в результате исследований.

1. На первом шаге в один класс объединяются те объекты, между которыми расстояние является минимальным.

2. На втором шаге производится пересчет матрицы расстояний с учетом вновь образованного класса.

Далее чередование пунктов 1 и 2 производится до тех пор, пока все объекты не будут объединены в один класс. Графическое представление результатов обычно осуществляется в виде дерева иерархической кластеризации. По оси X располагаются классифицируемые объекты (на одинаковом расстоянии друг от друга); по оси Y – расстояния, на основании

которых происходит объединение объектов в кластеры. Для определения «естественного» числа кластеров применяется критерий разбиения на классы в виде отношения средних внутрикластерных расстояний к межкластерным расстояниям. Глобальный минимум соответствует «естественному» числу классов, а локальные минимумы – *под-* и *над-* структурам (нижним и верхним границам).

Методы иерархического кластерного анализа различаются также по стратегии объединения (стратегии пересчета расстояний). Однако в стандартных статистических пакетах, к сожалению, не проводится оценка разбиения на классы, поэтому данный метод используется как предварительный с целью определения числа классов (обычно по соотношению межкластерных и внутрикластерных расстояний). Далее используется либо метод k-means, либо дискриминантный анализ, либо авторы, самостоятельно используя различные методы, доказывают отделимость классов.

При объединении *i*-го и *j*-го классов в класс *k*, расстояние между новым классом *k* и любым другим классом *h* пересчитывается одним из приведенных ниже способов (стратегии объединения). Расстояния между другими классами сохраняются неизменными. Наиболее распространенными являются следующие стратегии объединения (название несколько не соответствует содержанию; в соответствии с выбранными формулами производится пересчет расстояния от объектов до вновь образованного класса):

1. Стратегия «ближайшего соседа» – сужает пространство (классы объединяются по ближайшей границе)

$$d_{hk} = 1/2\, d_{hi} + 1/2 d_{hj} - 1/2 |d_{hi} - d_{hj}|.$$

2. Стратегия «дальнего соседа» – растягивает пространство (классы объединяются по дальней границе):

$$d_{hk} = 1/2\, d_{hi} + 1/2 d_{hj} + 1/2 |d_{hi} - d_{hj}|.$$

3. Стратегия «группового среднего» – не изменяет пространство (объекты объединяются в соответствии с расстоянием до центра класса) :

$$d_{hk} = (n_i/n_k)\, d_{hi} + (n_j/n_k)\, d_{hj},$$

где n_i, n_j, n_k – число объектов соответственно в классах *i,j,k*.

Первые две стратегии изменяют пространство (сужают и растягивают), а последняя его не изменяет. Поэтому, если не удается получить достаточно хорошего разбиения на классы с помощью третьей стратегии, а их все же необходимо выделить, то используются первые две, причем первая стратегия объединяет классы по ближайшим границам, а вторая – по дальним.

Таким образом, обычно в стандартных ситуациях используется стратегия «группового среднего». Если исследуемая группа достаточна разнородна, т.е. испытуемые, входящие в нее, значимо отличаются друг от друга по множеству характеристик, однако среди них необходимо выделить группы более сходные по всему профилю характеристик, то используется стратегия «дальнего соседа» (сужающая пространство). Если же группа достаточно гомогенна, тогда для выделения подгрупп среди очень схожих по характеристикам испытуемых следует использовать стратегию «дальнего соседа».

Рассмотрим фрагмент результатов исследования успешности деятельности команды – малой группы, ориентированной на решение деловой задачи и состоящей из молодых специалистов (инженеров-программистов), коллективно принимающих решение, выполняющих сложные работы в различном составе. Задача состоит в исследовании структуры данной команды и качественном описании характеристик каждой подгруппы. В качестве

характеристик были рассмотрены: зависимость от групповых стандартов, ответственность, работоспособность, трудовая активность, понимание цели, организованность, мотивация. Матрица смешения для 9 сотрудников приведена ниже.

Таблица 1. Матрица смешения для коллектива из 9 человек

№	Завис. от групп. стандарт.	Ответст.	Труд. актив.	Работо-способн.	Понимание цели	Мотивация
1	2.0	7.0	9.0	8.0	10.0	3.0
2	4.0	2.0	8.0	8.0	8.0	1.0
3	2.0	3.0	9.0	7.0	8.0	1.0
4	7.0	3.0	5.0	6.0	4.0	0.0
5	2.0	2.0	5.0	3.0	7.0	2.0
6	4.0	3.0	5.0	5.0	5.0	2.0
7	5.0	4.0	4.0	5.0	5.0	3.0
8	6.0	1.0	4.0	4.0	7.0	0.0
9	5.0	3.0	3.0	5.0	4.0	2.0

Используя метрику Евклида, получаем симметричную матрицу расстояний, которая является основой для кластерного анализа.

Таблица 2. Матрица расстояний, полученная с использованием метрики Евклида

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
C_1	0.00	6.16	5.00	10.3	8.72	8.43	8.77	10.5	10.3
C_2	6.20	0.00	2.65	6.30	6.32	5.39	6.56	6.20	7.30
C_3	5.00	2.65	0.00	7.70	5.92	5.83	7.21	7.50	8.10
C_4	10.3	6.32	7.68	0.00	6.93	3.87	4.12	4.40	3.60
C_5	8.70	6.32	5.92	6.90	0.00	3.61	4.80	4.80	5.20
C_6	8.40	5.39	5.83	3.90	3.61	0.00	2.00	4.20	2.40
C_7	8.80	6.56	7.21	4.10	4.80	2.00	0.00	4.90	2.00
C_8	10.5	6.24	7.48	4.40	4.80	4.24	4.90	0.00	4.50
C_9	10.3	7.28	8.12	3.60	5.20	2.45	2.00	4.50	0.00

Результат применения агломеративного иерархического метода КА к полученной матрице при использовании пакета STATISTICA – дерево классификации – представлен на рис.1.: по горизонтальной оси откладываются на одинаковом расстоянии номера объектов (членов команды), по вертикальной оси – расстояние, на котором объединяются эти объекты.

Можно заметить, что выделилось два класса: в один вошли объекты 5, 8, 9, 7, 6, 4, а в другой – 3, 2, 1. Отделимость классов оценивается сравнением внутрикластерных и межкластерных расстояний на качественном уровне.

Примененный к результатам эмпирических исследований агломеративный иерархический метод КА позволяет выделить «естественное» число классов, а также *под-* и *над-*структуры. Он будет более эффективным при использовании оценок разбиения на классы.

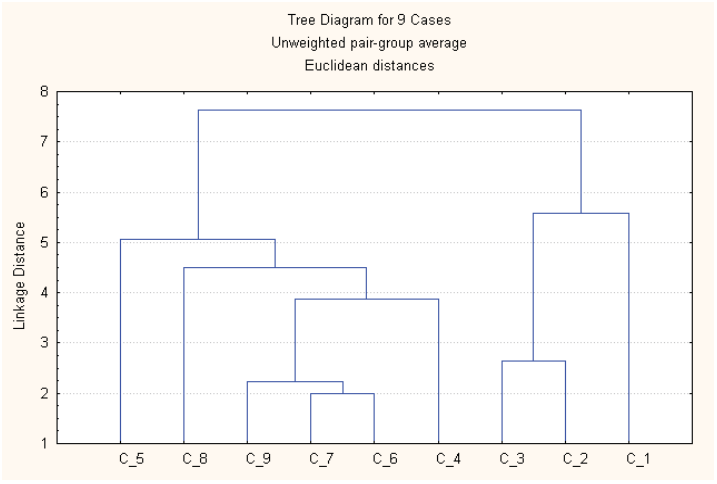

Рис. 1. Дерево классификации

Для определения «естественного» числа кластеров, на которые может быть разбита совокупность объектов, а возможно, и для выделения более «тонкой» структуры применялся следующий критерий: на каждом уровне иерархической кластеризации выполнялось разбиение множества на данное число классов. В основу примененной для этого формулы была заложена идея физической плотности или, точнее, объема пространства, занимаемого данным множеством объектов (Савченко, Рассказова, 1989). Для каждой пары кластеров оценивалась степень их внутренней связанности друг с другом. Для этого вычислялось среднее внутрикластерное расстояние для каждого кластера из заданной пары; если при этом в класс входит всего один элемент, то расстояние соответствует минимальному расстоянию до какого-либо из элементов. Если в классе более одного элемента, но все различия между ними равны 0, то в формуле отражается аналогия с объемом пространства, занимаемого одним объектом. Формула учитывает, что в данном случае в одной точке пространства находится лишь один объект с большей «удельной плотностью».

В качестве оценки связанности берется отношение среднего внутрикластерного расстояния к межкластерному:

$$\pi = \frac{a_i + a_j}{2b_{ij}},$$

где a_i и a_j – средние внутрикластерные расстояния классов i и j;
b_{ij} – среднее межкластерное расстояние между этими же классами.

Оценка «естественного» разбиения производится по следующей формуле:

$$S = \frac{1}{k}\sum_{i=1}^{k} \max_j \pi_{ij}$$

Отметим некоторые свойства такого разбиения: если все различия между объектами равны между собой, то S для такого случая равна 1; разбиения, получаемые с помощью вышеописанного алгоритма, имеют оценку не более 1. Итак, будем считать значение критерия такого разбиения, когда все объекты объединены в один кластер, равным 1.

Минимум значения функции *S* определяет наилучшее разбиение множества объектов на кластеры. Изображение на одном графике дерева кластеризации и значений функции *S* позволяет выявлять не только оптимальное разбиение, но и *под-* и *над-* структуры, которые соответствуют локальным минимумам функции S и позволяют обнаружить в множестве разные уровни объединения. Таким образом, описанный метод кластерного анализа позволяет выявлять иерархическую организацию множества объектов, используя только матрицу различий между ними.

Однако в стандартных пакетах, как отмечалось выше, такая оценка, к сожалению, не предусмотрена. Для получения более детальной информации о полученных классах используются другие методы кластеризации: например, дендритный анализ дает возможность проследить близость объектов в классах и более подробно изучить их структуру; метод k-means позволяет качественно описать каждый класс объектов и провести сравнительный анализ степени выраженности исследуемых характеристик у представителей обоих классов.

При анализе данных социально-психологических исследований взаимоотношений в коллективах помимо разбиения на классы необходимо решить вопрос о том, какие именно объекты (характеристики, признаки) связывают классы друг с другом. В этом случае целесообразным является использование *дендритного метода кластерного анализа*, который часто применяется совместно с иерархическим. Дендрит в данном случае – это ломаная линия, которая не содержит замкнутых ломаных и в то же время соединяет любые два элемента. Он определяется не единственным способом, поэтому предлагается построение дендрита, у которого сумма длин связей минимальна.

Итак, объекты – это вершины дендрита, а расстояния между ними – дуги. На первом этапе к каждому объекту находится ближайший (находящийся к нему на минимальном расстоянии) объект и составляются пары. Число пар равно числу объектов. Далее, если есть симметричные пары (например: *i*_____*j*, *j*_____*i*), то одна из них убирается; если в двух парах присутствует один и тот же элемент, то пары объединяются через этот элемент. Например, две пары:

*i*_____*j*,
*j*_____*k*

объединяются в связку *i*_____*j*_____*k* .

На этом заканчивается построение скоплений (плеяд) первого порядка. Затем определяются минимальные расстояния между объектами скоплений первого порядка, и эти скопления объединяются до тех пор, пока не будет построен дендрит. Группы объектов считаются вполне отделимыми, если длина дуги между ними $d_{lk} > C_p$,

где $C_p = C_{cp} + S$, C_{cp} – средняя длина дуги, S – стандартное отклонение.

Дендриты могут принимать форму розетки, амебообразного следа, цепочки. При совместном использовании иерархического КА и метода дендрита распределение элементов по классам получают, применив КА, а взаимосвязи между элементами анализируются с помощью дендрита.

Применение дендритного анализа к рассматриваемым данным позволило получить следующий дендрит (см. рис. 2).

Итак, в описанном выше случае C_p=4.8. Это означает, что выделяются три класса, что несколько отличается от результата, полученного с помощью агломеративного метода. Из первого класса, в который входили объекты 1, 3, 2, отделился первый член коллектива. Во второй класс вошли объекты 8, 4, 9, 7, 6, 5 (аналогично результатам, полученным с помощью агломеративного метода).

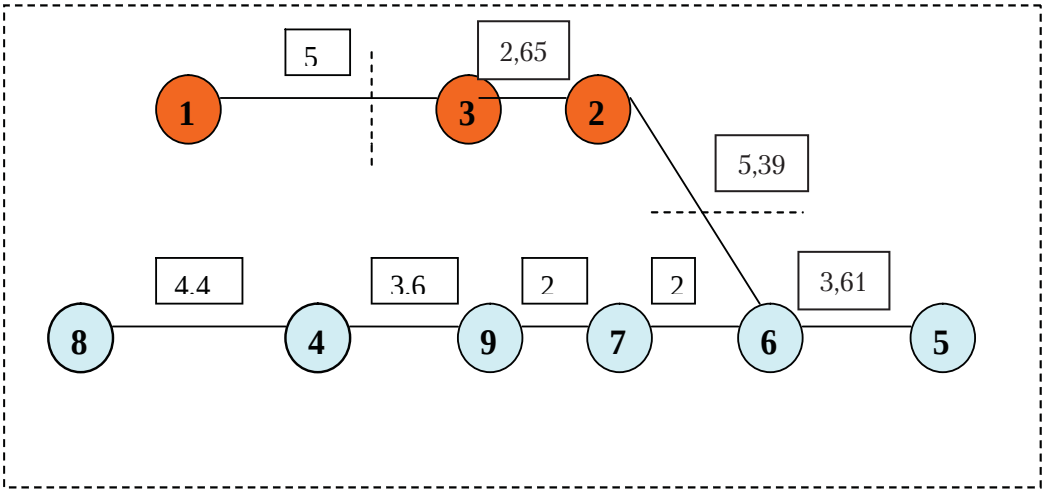

Рис. 2. Дендрит (форма простого дерева): над дугами дендрита указаны расстояния между объектами

Применение такого метода позволяет получить дополнительную информацию о том, какие объекты связывают классы друг с другом. В нашем случае это 2 и 6 объекты (члены коллектива). Данная структура аналогична социометрической, однако получена она на основе результатов тестирования. Дальнейший анализ дендрита позволит выделить группы совместимых людей (которые наиболее эффективно решают поставленные задачи в ходе совместной деятельности) либо выявить тех, кто лучше работает в одиночку, например, объект 1; 8 объект находится на границе отделимости, поэтому, возможно, ему лучше давать индивидуальные задания.

Помимо агломеративных иерархических методов существует также большое количество *итеративных методов кластерного анализа*. Основное отличие их состоит в том, что процесс классификации начинается с задания начальных условий: это может быть число классов, критерий завершения классификации и т.д. К таким методам относятся, например, дивизивные методы, методы k-means и другие, требующие от исследователя интуиции и творческого подхода. Еще до проведения классификации необходимо представлять, какое количество классов должно быть образовано, когда закончить процесс классификации и т.д. От верно выбранных начальных условий будет зависеть результат классификации, поскольку некорректно выбранные условия могут приводить к «размытости» классов. Таким образом, эти методы используются, если есть теоретическое обоснование, например, количества ожидаемых классов, а также после проведения иерархических методов классификации, которые позволяют выработать наиболее оптимальную стратегию исследования.

Метод k-means можно отнести к итеративным методам эталонного типа. Название ему было дано Дж.Мак-Куином. Существует много различных модификаций данного метода. Рассмотрим одну из них.

Пусть в результате проведенного исследования получена матрица измерений n объектов по m характеристикам. Множество объектов необходимо разбить на k классов по всем исследуемым характеристикам.

На первом шаге из n объектов выбираются k точек либо случайным образом, либо **ис**ходя из теоретических предпосылок. Это и есть эталоны. Каждому из них присваивается порядковый номер (номер класса) и вес, равный единице.

На втором шаге из оставшихся n-k объектов извлекается один и проверяется, к какому из классов он ближе, для чего используется одна из метрик (к сожалению, в основных статистических пакетах используется только метрика Евклида). Рассматриваемый объект относится к тому классу, к эталону которого он наиболее близок. Если есть два одинаковых минимальных расстояния, то объект присоединяется к классу с минимальным номером.

Производится перерасчет эталона, к которому присоединен новый объект, и его вес возрастает на единицу.

Пусть эталоны представлены таким образом:

$$e(1,0)=(x_{11,}x_{12},...,x_{1i},...,x_{1m})$$
$$.....$$
$$e(k,0)=(x_{k1,}x_{k2},...,x_{ki},...,x_{km})$$

Тогда если рассматриваемый объект j относится к эталону k, то данный эталон (т.е. центр образовавшегося класса) пересчитывается следующим образом:

$$e(j,1) = (\frac{x_{j1}+x_{k1}\bullet v_{j0}}{2};...;\frac{x_{jn}+x_{kn}\bullet v_{j0}}{2})$$

здесь v_{j0} – вес эталона j в нулевой итерации.

Остальные эталоны остаются неизменными.

Далее процедура повторяется до тех пор, пока все n-k объекты не будут отнесены к каким-либо эталонам. Веса эталонов накапливаются.

Чтобы получить устойчивое разбиение, новые эталоны после разнесения всех объектов принимаются за начальные, и далее процедура повторяется с первого шага. Веса классов продолжают накапливаться. Новое распределение по классам сравнивается с предыдущим, если различие не превышает заданного уровня, т.е. распределения можно считать не изменившимися, то процедура классификации заканчивается.

Существует две модификации данного метода. В первой пересчет центра кластера происходит после каждого присоединения, во второй – в конце отнесения всех объектов к классам; минимизация внутрикластерной дисперсии осуществляется в большинстве итерационных методов кластерного анализа.

Обычно в методе k-means реализуется процедура построения усредненных профилей каждого класса (см. рис. 3), что дает возможность проводить качественный анализ выраженности признаков у представителей каждого класса. Для сравнения классов по выраженности тех или иных характеристик используется процедура, подобная ANOVA, сравнивающая внутрикластерные и межкластерные дисперсии по каждой характеристике и тем самым позволяющая осуществить проверку значимости различия классов по исследуемым характеристикам.

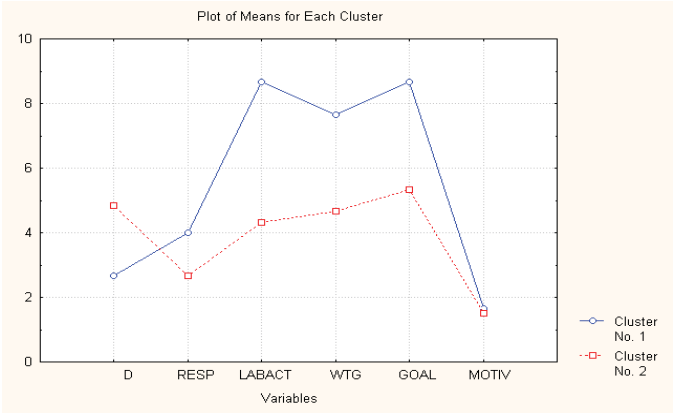

Рис. 3. Усредненные профили классов

Таблица 3. Номера объектов и расстояния от центра классов

Номера объектов первого класса

	1	3	2
Distance	1.484488	1.097134	0.693889

Номера объектов второго класса

	8	4	9	7	6	5
Distance	1.357421	1.566430	0.535758	0.855267	1.272938	0.822147

Анализ профилей показывает, что в первый класс (табл. 3) попали члены коллектива, характеризующиеся незначительной зависимостью от группы, средним уровнем ответственности и высокой трудовой активностью, работоспособностью, пониманием цели. Во вторую группу (более многочисленную) вошли сотрудники, характеризующиеся значительной зависимостью от групповых стандартов, низким уровнем ответственности, трудовой активности, работоспособности и понимания общей цели. На тех, кто вошел в состав первой группы, может быть возложена ответственность, они могут самостоятельно принимать решения и т.д.; вторая группа – это исполнители, за выполнением порученных заданий которыми необходим постоянный контроль. Заметим лишь, что мотивация низкая у обеих групп, что связано, возможно, с невысокой оплатой труда. В табл. 4 представлены результаты сравнительного анализа, демонстрирующие значимые отличия классов по трем характеристикам: трудовая активность, работоспособность и понимание цели.

Таблица 4. Анализ отделимости классов (жирным шрифтом выделены те характеристики, по которым наблюдается значимое различие между классами).

	Межкл. расст.	**Внутрикл. расст.**	**Степ.свободы**	**F**	**p**
D	9.38	17.50	7	3.75	0.093821
RESP	3.55	19.33	7	1.28	0.293890
LABACT	**37.55**	**4.00**	**7**	**65.72**	**0.000084**
WTG	**18.00**	**6.00**	**7**	**21.00**	**0.002536**
GOAL	**22.22**	**12.00**	**7**	**12.96**	**0.008735**
MOTIV	0.055	10.16	7	0.038	0.850495

К оригинальным методам, в основе которых лежит психологическая теория, можно отнести *кластерный анализ на основе теории Выготского*. В работе «Мышление и речь» Выготский описывает различные генетические ступени развития понятий. В частности, он выделяет в качестве одного из важнейших этап образования комплексов, являющихся прообразами научных понятий. Он пишет, что в основе комплекса лежат фактические связи между объектами, устанавливаемые в непосредственном опыте. Поэтому такой комплекс представляет собой прежде всего конкретное объединение предметов на основании их фактической близости друг с другом. Далее он выделяет пять форм комплексов, а именно: ассоциативный комплекс, комплекс-коллекция, цепной комплекс, диффузный комплекс, псевдопонятия. Важно сразу же отметить, что во всех типах комплексов возможны любые ассоциативные связи, причем их характер может быть совершенно различным между различными парами элементов, участвующих в образовании одного и того же комплекса. Так что важнейшей особенностью образования комплексов является множественность типов ассоциативных связей между элементами, объединяемыми в комплекс. Заметим, что в качестве частного случая различий между элементами может выступать различие по какому-либо критерию. В кластерном анализе таким критерием является (моделируется) расстояние. Поскольку характер связей в ассоциативном комплексе может быть различным, то формализация осуществляется через задание на одном и том же множестве элементов нескольких различных типов попарных расстояний (или различий) между ними.

Допустим, что в описанном нами примере предметом изучения являются отношения между членами некоей малой группы, например, производственной, научной или учебной. Для одной и той же группы может быть выделено несколько типов отношений: производственные, личные, общность увлечений и т.д. Тогда для какой-либо из групп экспериментально определяется структура отношений каждого типа и строится матрица попарных расстояний (или близости) между членами группы по каждому типу отношений.

Формальное описание ситуации сводится к следующему. Задано множество M элементов $A_1, A_2,...A_n$ и множество типов попарной близости этих элементов. Пусть количество этих типов m. Различные типы близости отличаются друг от друга тем, что каждый представляет собой близость по какому-либо качеству, присущему всем элементам множества. Таким образом, выделяются m качеств каждого элемента и производится сравнение (вычисление расстояний или различий) по каждому из этих качеств, что и дает m типов близости элементов. Для каждого типа близости задается матрица попарных расстояний (или различий), отражающая структуру множества элементов m по отношению к данному типу близости. Всего должно быть задано m таких матриц.

Покажем теперь, как в рамках данной формальной схемы могут быть описаны алгоритмы образования комплексов различных типов.

1. *Ассоциативный кластер*. Согласно Выготскому, в ассоциативном комплексе прежде всего выделяется элемент, который будет образовывать его ядро, затем остальные элементы объединяются с ядром. И здесь Выготский отмечает следующую характерную особенность данного комплекса: «Элементы могут быть вовсе не объединены между собой. Единственным принципом их обобщения является их фактическое родство с основным ядром комплекса. Связь, объединяющая их с этим последним, может быть любой ассоциативной связью» (Выготский, 1982, с. 142).

Дадим описание простейшего варианта алгоритма образования ассоциативного кластера в терминах приведенной выше формальной схемы. Сначала из заданного множества M элемен-

тов выбирается один, который будет играть роль ядра ассоциативного кластера. Ясно, что можно построить столько ассоциативных кластеров, сколько элементов в множестве *M*, выбирая поочередно в качестве ядра все элементы множества. Итак, выберем один элемент A_k. Далее, по каждому качеству (т.е. для каждой матрицы расстояний) выбирается элемент, ближайший к элементу A_k. Таким образом, мы получаем *m* или более элементов, если по каким-либо признакам выделяются два или более элементов, отстоящих от A_k на одно и то же минимальное по этому признаку расстояние. Совокупность элемента A_k как ядра и всех таким образом выбранных ближайших к нему элементов по каждому признаку и составляет ассоциативный кластер.

Возможны и более сложные алгоритмы, например, если с самого начала в качестве ядра ассоциативного кластера выбирать не один элемент, а несколько. Такой вариант кластерного анализа мы будем называть обобщенным ассоциативным кластером. Опишем алгоритм его образования более подробно.

Сначала выбирается множество элементов, которые в совокупности будут составлять ядро обобщенного ассоциативного кластера. Далее по каждому признаку для каждого из элементов ядра отбираются ближайшие по выбранному признаку элементы, а величины этих минимальных расстояний фиксируются. Затем из всех расстояний выбирается наименьшее, и происходит отбор только тех элементов, которые находятся на минимальном расстоянии от какого-либо из элементов ядра. Эта процедура повторяется для всех качеств. При этом в переборе элементов, естественно, не участвуют те, что составляют ядро кластера. Совокупность элементов ядра и всех элементов, выбранных в соответствии с описанной процедурой, и является обобщенным ассоциативным кластером. Элементы ассоциативного комплекса (по Выготскому) могут вовсе не быть объединены между собой, а находиться в ассоциативной связи лишь с ядром комплекса. Это означает, что a priori могут быть заданы не все расстояния, т.е. множество элементов упорядочится лишь частично.

Рассмотрим конкретный пример применения простейшего алгоритма образования ассоциативного кластера для анализа отношений в малой группе.

Количество членов малой группы, т.е. элементов рассматриваемого множества, *n*=9. Было выбрано *m*=3 различных типов отношений между членами малой группы: 1) взаимоотношения, связанные с основной работой, 2) взаимоотношения, связанные с неделовыми формами общения, 3) взаимоотношения, связанные с участием в дополнительной работе. По каждому типу отношений методами экспертных оценок были получены матрицы попарных различий (расстояний) между всеми членами группы.

В соответствии с описанным выше простейшим алгоритмом образования ассоциативного кластера были построены все 9 кластеров, причем в качестве ядра были выбраны поочередно все члены малой группы. На рис. 4 представлен пример полученного ассоциативного кластера, в котором в качестве ядра взят элемент A_1.

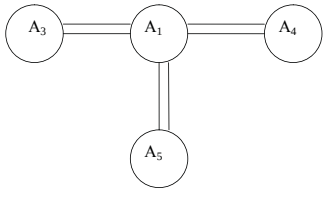

Рис. 4. Ассоциативный кластер с ядром A_1

2. *Цепной кластер.* «Цепной комплекс строится по принципу динамического временного объединения отдельных звеньев в единую цепь и переноса значения через отдельные звенья этой цепи. Каждое звено соединено... с предшествующим... (и)... последующим, причем самое важное отличие этого типа комплекса в том, что характер связи или способ соединения одного и того же звена с предшествующим и последующим может быть совершенно различным» (Выготский, 1982, с. 144).

Теперь приведем описание алгоритма образования цепного кластера в принятых нами терминах формальной модели. Сначала из заданного множества *m* элементов выбирается один, который станет первым элементом, составляющим цепной кластер. Затем для каждого качества (т.е. для каждой матрицы расстояний из *m* заданных матриц) выбирается элемент, ближайший к первому. Из полученных *M* минимальных расстояний выбирается наименьшее и фиксируется номер соответствующей матрицы и номер элемента – этот элемент и будет вторым в цепном кластере. Далее процедура повторяется для второго элемента, причем первый из процесса отбора исключается. Процесс повторяется столько раз, сколько элементов в множестве *M*.

Заметим, что если на каком-либо шаге построения цепного кластера минимальная величина будет не у одной, а у двух или более пар элементов, то в этом случае может быть построено несколько эквивалентных цепных кластеров. Графическое изображение построенного нами цепного кластера, начинающегося с элемента A_1, представлено на рис. 5, где видно, как к группе из элементов A_1, A_3, A_4 присоединяются последовательно остальные элементы. Однако необходимо подчеркнуть, что в данном исследовании цепной кластер менее информативен, чем ассоциативный, тем не менее он предоставляет дополнительные к ассоциативному кластеру сведения.

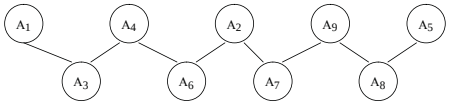

Рис. 5. Цепной кластер с ядром A_1.

3. *Ассоциативно-цепной кластер.* Как уже отмечалось, процедуры построения ассоциативного и цепного кластеров решают различные содержательные задачи: ассоциативный выявляет все элементы, ближайшие к ядру по различным свойствам, а цепной показывает связь данного начального элемента последовательно со всеми остальными элементами множества. Представляется целесообразным разработать такой алгоритм, который обладал бы преимуществами как ассоциативного, так и цепного кластера. Далее приведем описание одного из возможных вариантов построения ассоциативно-цепного кластера.

Выберем сначала один элемент, который будет ядром ассоциативно-цепного кластера, в этом качестве может выступать любой элемент множества. Затем применим алгоритм образования простейшего ассоциативного кластера. Рассмотрим далее множество элементов, составивших простейший кластер. Применим к этому множеству элементов алгоритм построения обобщенного ассоциативного кластера. Далее к получившемуся множеству элементов, которые составляют обобщенный кластер, снова применим алгоритм образования. Будем повторять эту процедуру до тех пор, пока в строящийся кластер не

объединятся все элементы исходного множества. Полученную в результате описанного процесса структуру и будем называть ассоциативно-цепным кластером. Это название оправданно тем, что структура подобного кластера представляет собой центральный простейший ассоциативный кластер и цепочки из элементов, составляющих простейший кластер. На рис. 6 представлен пример построения ассоциативно-цепного кластера для рассматриваемых нами экспериментальных данных. В качестве исходного элемента взят элемент A_1.

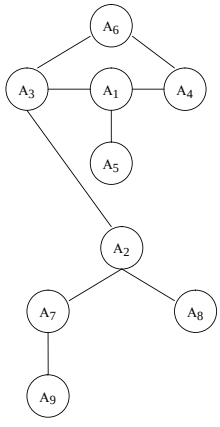

Рис. 6. Ассоциативно-цепной кластер с ядром A_1

Мы видим, что к образовавшемуся простейшему ассоциативному кластеру с ядром A_1 присоединяются элементы A_2, A_6, A_7 и, наконец, элементы A_8 и A_9 на различных итерациях. Если коротко охарактеризовать смысл ассоциативно-цепного кластера, то можно сказать, что он описывает структуру заданного множества элементов по отношению к одному выделенному (на рис. 6 это элемент A_1).

4. *Кластер-коллекция.* Рассмотрим, наконец, тип кластера, соответствующий комплексу-коллекции Выготского. Характеризуя его, ученый пишет, что комплексы этого типа «больше всего напоминают то, что принято называть коллекциями. Здесь различные неконкретные предметы объединяются на основе взаимного дополнения по какому-либо одному признаку и образуют единое целое, состоящее из разнородных, взаимно дополняющих друг друга частей». И далее: «Эта форма мышления часто соединяется с описанной выше ассоциативной формой. Тогда получается коллекция, составленная на основе различных признаков» (Выготский, 1982, с. 142–143).

Рассмотрим теперь описание простейшего варианта алгоритма образования кластера-коллекции в терминах приведенной выше формальной модели. Заметим, что в результате применения алгоритма построения кластера-коллекции мы должны получить набор элементов, отличающихся друг от друга хотя бы по одному признаку. К такому результату приводит, например, следующий алгоритм: сначала задается некоторый порог различия (или расстояния), при котором два элемента с разницей больше выбранного порога считаются различными. Очевидно, что результат (кластер-коллекция) будет зависеть от величины порога.

Далее раздельно для каждого признака (т.е. для каждой матрицы расстояний) применяется обычный метод кластерного анализа. По каждому признаку на основе результатов обычного анализа выбирается такое деление на кластеры, при котором расстояния между ними превышают заданный порог.

Затем рассматриваются одновременно все разбиения, выполненные по различным свойствам, и фиксируются все пересечения и разности множеств элементов, составляющих эти кластеры. Очевидно, что множества элементов, полученные таким способом, обладают следующим свойством: элементы двух различных множеств находятся хотя бы по одному признаку на расстоянии, превышающем выбранный порог. Если теперь возьмем по одному (любому) элементу из всех полученных множеств, то получим кластер-коллекцию.

Рассмотрим пример построения кластера-коллекции для наших экспериментальных данных. Напомним, что множество состоит из 9 элементов и имеются три матрицы попарных расстояний между ними. Пусть величина порога будет h=7. Проведя обычный кластерный анализ для каждой из трех матриц расстояний и применив описанную выше процедуру при величине порога h=7, получим следующие разбиения.

Для первой матрицы – три кластера:

$$\{A_1 A_2 A_3 A_4 A_5 A_6 A_7\}\ \{A_8\}\ \{A_9\}.$$

Для второй – четыре кластера:

$$\{A_1 A_2 A_3 A_4 A_5 A_6\}\ \{A_7\}\ \{A_8\}\ \{A_9\}.$$

Для третьей – четыре кластера:

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2 A_7\}\ \{A_8\}\ \{A_9\}.$$

Выбирая в соответствии с описанной выше процедурой пересечения и разности всех полученных кластеров, получим в результате следующий набор множеств:

$$\{A_1 A_3 A_4 A_5 A_6\}\ \{A_2\}\ \{A_7\}\ \{A_8\}\ \{A_9\}$$

Таким образом, в кластер-коллекцию входят элементы A_2, A_7, A_8 A_9 и еще один (любой) элемент первого множества, например, A_1. Очевидно, что элементы кластера-коллекции A_1, A_2, A_7, A_8 A_9 отличаются друг от друга хотя бы по одному признаку на величину, большую h=7. Так, например, элементы A_1 и A_2 отличаются лишь по одному третьему признаку, элементы A_1 и A_7 по второму и третьему, а, скажем, элементы A_8 и A_9 – по всем трем.

Метод латентных классов

Цель создания моделей с латентными переменными состоит в объяснении наблюдаемых переменных и взаимосвязей между ними: при заданном значении наблюдаемых переменных конструируется множество латентных переменных и подходящая функция, которая достаточно хорошо аппроксимировала бы наблюдаемые переменные, а в конечном счете плотность вероятности наблюдаемой переменной.

В факторном анализе основной акцент делается на моделирование значений наблюдаемых переменных из корреляций и ковариаций, а в методах латентно-структурного анализа – на моделирование распределения вероятности наблюдаемых переменных.

Метод латентных классов можно использовать для дихотомических переменных и порядковых шкал. Наблюдаемые переменные могут быть измерены в дихотомической шкале наименований, т.е. являются переменными (0,1) (x_i =1 – наличие признака и x_i=0 – отсутствие признака). Тогда наблюдаемые вероятности могут быть объяснены с помощью латентных переменных, т.е. с помощью латентных распределений и соответствующих условных распределений. (Лазарфельд, 1996).

Объясняющее уравнение первого рода имеет вид:

$$\rho_i = \int_{-\infty}^{+\infty} h_i(x_i \mid \varphi) g(\varphi) d\varphi$$

где наблюдаемые переменные – x_i; плотность вероятности наблюдаемых переменных – ρ_i; множество латентных переменных – φ, плотность вероятности латентных переменных – $g(\varphi)$.

Объясняющее уравнение n-го порядка имеет вид:

$$\hat{\rho}_n = \int_{+\infty}^{-\infty} h_i(x_i|\varphi)h_k(x_k|\varphi)...h_n(x_n|\varphi)g(\varphi)d\varphi$$

Основным предположением всех моделей латентных структур является локальная независимость. Это следует понимать так: для данной латентной характеристики наблюдаемые переменные независимы в смысле теории вероятностей. Аксиома локальной независимости имеет вид:

$$h(\frac{\rho}{x}|\varphi) = \prod_{i=1}^{n} h_i(x_i|\varphi)\frac{\rho}{x} = (x_1,...,x_n)$$

Условная вероятность $h_i(x_i\varphi)$ называется операционной характеристикой вопроса, т.е. это вероятность получения правильной оценки того, что наблюдаемый признак j имеет место, если его латентная характеристика известна. Если φ непрерывна, то операционная характеристика называется характеристикой кривой, или следом.

По дискретности или непрерывности и по виду характеристической кривой различают следующие модели: модели латентных групп (латентную вероятность р группы можно обозначить через g, а операционную характеристику – через $h_i(\varphi_p) = h_i(x_i|\varphi_p)$); модель латентных профилей (обобщение модели латентных групп, когда наблюдаемые переменные считаются непрерывными); модель латентных расстояний, которая имеет в качестве характерной кривой функцию скачка.

Рассмотрим одну из моделей латентных групп (дискретная латентная характеристика). На основе модели Роста нами был реализован метод латентно-структурного анализа, или модель латентных классов для нормального распределения данных. Таким образом, решается следующая задача: по матрице ответов испытуемых на вопросы какого-либо теста структурируется само множество испытуемых по близости (похожести) профилей ответов.

Для этой цели сначала произвольно задаются два параметра, которые являются скрытыми – латентными, так как истинное их значение предстоит определить в процессе работы метода. Это :

1. Относительное число испытуемых в классе (мы задавали его первоначально $P(k) = 1/k$).

2. Характеристический параметр класса $r(i, k)$ – матрица вероятности появления определенного ответа на i-й вопрос, если испытуемый относится к k-му классу. Он должен быть различным для разных классов. Мы задавали его и одинаковым для испытуемых, принадлежащих к одному классу, и различным для каждого класса. Предполагается, что условная вероятность такого события, как ответ испытуемого категории q на j вопрос, постоянна для всех испытуемых, принадлежащих к классу k. Вероятность появления ответа категории $q(1,2,...,Q)$ равна вероятности q, являющейся суммой реализаций дихотомической случайной переменной.

В конце определяются для априорно заданного числа классов истинное относительное число испытуемых в классах и истинный параметр, определяющий вероятность появления определенного ответа на i-й вопрос, если испытуемый относится к k-му классу, что отражается в профилях, характеризующих именно данную группу испытуемых.

Мы вычисляли также наиболее вероятный профиль ответов испытуемых, принадлежащих к данному классу. Структура данных включает:

1. Матрицу профилей ответов.

2. Матрицу априорных вероятностей: вероятности определенного ответа на i-й вопрос при условии, что испытуемый относится к *k*-му классу.

3. Относительное число испытуемых в классе.

В основе модели лежит формула Байеса, которая связывает априорную вероятность с апостериорной. Общая методология сводится к введению априорной плотности распределения параметров и последующему нахождению по формуле Байеса их апостериорной плотности распределения (с учетом экспериментальных данных).

Априорные распределения могут задаваться (1) стандартным способом (априорная вероятность пропорциональна числу классов); (2) исходя из профессиональных соображений, т.е. априорно задаются две латентные характеристики:

1. Количество латентных классов (*k*) и соответствующее им относительное число испытуемых в классе *P*(*k*);

2. Параметр, определяющий вероятность определенного ответа на 1-й вопрос при условии, что испытуемый относится к *k*-му классу *r*(*k*).

Вероятность появления 1-го паттерна профиля :

$$P_i = \sum_{k=1}^{k} P(k) r_i \quad .$$

Далее по формуле Байеса вычисляется апостериорная вероятность:

$$P(k \mid i) = \frac{P(k) r_i(k)}{P(i)} \quad .$$

Алгоритм метода латентных групп.

Задаем:

а) количество латентных классов *K*,

б) количество вопросов *M*,

в) количество возможных категорий ответов *Q*,

г) количество испытуемых *N*,

д) начальное распределение.

P(*k*) – относительное число испытуемых, которые входят в класс, например *P*(*k*) = 1/*k*:

Задаем начальные значения характеристик параметров классов *r*(*i*,*k*) ; *k* = 1,..., *k*; *i*=1,..., *M; r*(*i*,*k*) – параметр, определяющий вероятность появления определенного ответа на *i*-й вопрос, если испытуемый относится к *k*-му классу.

Вводим *Xij* – ответ *i*-го испытуемого на *j*-й вопрос: *i*=1,...,*N*; *j*=1,...,*M*.

Определяем множество различных паттернов ответов:

$\overline{a_i}$, где $x_{ij} = a_{ij}$, a_{ij} – ответ на *j*-й вопрос.

Считаем количество таких паттернов: *n*(*i*), *i*=1,...,L; *n*($\overline{a_i}$).

Вычисляем вероятность появления паттерна $\overline{a_i}$ при условии, что он генерируется испытуемым, относящимся к *k*-му классу:

$$\rho_k(\bar{a}_i) = \prod_{j=1}^{M} \left(\frac{Q!}{a_{ij}(Q-a_{ij})!} r^{a_{ij}}(j,k)(1-r(j,k))^{Q-a_{ij}}\right) .$$

Вычисляем вероятность появления такого паттерна:

$$\rho(\bar{a}_i) = \sum_{k=1}^{k} \rho_k(\bar{a}_i) P(k) .$$

Вычисляем апостериорную вероятность того, что испытуемый относится к классу k, если он ответил $\bar{a}_i$:

$$\rho_{\bar{a}_i(k)} = \frac{\rho_k(\bar{a}_i)}{\rho(\bar{a}_i)} .$$

Вычисляем математическое ожидание количества паттернов у испытуемых класса k:

$$E(\bar{a}_i, k) = n(\bar{a}_i)\rho_{\bar{a}_i}(k) .$$

Считаем оценку относительного числа испытуемых, относящихся к классу k:

$$\hat{P}(k) = \frac{1}{N}\sum_{\bar{a}_i} E(\bar{a}_i, k) .$$

Вычисляем математическое ожидание количества паттернов, в которых ответ на j-й вопрос есть $x \in \{0,1,...,Q\}$, при условии, что отвечающие относятся к классу k:

$$E_x(j,k) = \sum_{a_{ij}=x} \frac{E(\bar{a}_i, k)}{\hat{P}(k)} \quad ; \text{x = 0,1,...,Q.}$$

Вычисляем оценку параметров:

$$\hat{r}(j,k) = \frac{1}{N}\frac{1}{Q}\sum_{x=0}^{Q} xE_x(j,k) .$$

Если $\max_{j,k} |r(j,k) - r(j,k)| \le \delta$, то мы получаем интересующие нас параметры классов, т.е. $\hat{P}, \hat{r}(i,k) u \rho_{\bar{a}_i}(k)$

В противном случае процедура повторяется.

Также нами были разработаны четыре варианта оценки кластерных разбиений.

Есть множество испытуемых X. $\|X\|=N$ – мощность множества X равна N, т.е. N – испытуемых. В результате LSA мы получаем для каждого из K классов и N испытуемых:

P_i^k $_{1\le k\le K}$ – вероятность для i-го испытуемого принадлежать к k-му классу.

Определяя max Pi, мы относим испытуемого i к классу, к которому он принадлежит, с максимальной вероятностью.

Разбивая множество X на классы указанным выше образом, получаем:

X^k – множество испытуемых, попавших в k-й класс;

$||X^k||$ – количество испытуемых, попавших в k-й класс.

Тогда можно предложить следующие оценки разбиений: средняя «четкость» кластеров, наименьшая «четкость» кластеров, интегральная «четкость» кластеров, связность кластеров.

Аналогично методу иерархической кластеризации, описанному выше, наиболее верно отражающей реальную структуру оказалась оценка, названная нами – связность кластеров.

$$\pi_{uv} = \frac{2(\sum_{i \in x^u} \rho_i^v + \sum_{j \in x^v} \rho_i^u)}{||X^v|| \sum_{i \in X^u} \rho_i^u + ||X^u|| \sum_{j \in X^v} \rho_j^v} \qquad L_4 = \sum_{u=1}^{K} \max_v \pi_{uv} .$$

Тогда возьмем два класса; их параметры – относительное число испытуемых в классе, вероятность для i-го испытуемого принадлежать к k-му классу. Из двух вероятностей выбирается большая, что и определяет класс, к которому «принадлежит» испытуемый (реально испытуемый может не принадлежать ни к одному из классов). Если при этом в одном из анализируемых классов не оказалось ни одного испытуемого, то суммарная вероятность по этому классу равна 0.

Вызывает несомненный интерес тот факт, что именно «связность» работает в обоих методах, разработанных в лаборатории математической психологии, – методе латентных классов и методе иерархической кластеризации. При кластерном анализе это можно было оценить и визуально, изучая картинку дерева. В ЛСА это можно заметить следующим образом: до данного количества кластеров (определяемых этой оценкой) профили классов существенно отличаются друг от друга, а далее заметно лишь незначительное отличие.

Данный метод позволяет выделить наиболее типичные паттерны восприятия стимулов и проанализировать их профили. Метод основан на вероятностном подходе, поэтому является более универсальным по сравнению с другими методами кластерного анализа. Наиболее часто метод ЛСА используется при адаптации методик, так как позволяет выделить типичные паттерны ответов и в соответствии с ними структурировать множество испытуемых, а для каждого типа оценить апостериорную вероятность.

В представленной статье описаны различные методы кластерного анализа и показано, в каких случаях их можно применять с наибольшей эффективностью по отдельности, а также совместно друг с другом. Итак, в статье представлены стандартные методы, реализованные в наиболее часто используемых статистических пакетах, их развитие и усовершенствование, которое реализовано на данном этапе только в оригинальных пакетах, а также оригинальные методы, отсутствующие в статистических пакетах.

Литература

Выготский Л.С. Мышление и речь // Собр. соч.: В 6 т. Т. 2. М.: Педагогика, 1982.

Глинский В.В., Ионин В.Г. Статистический анализ данных, М.: Филин, 1998.

Головина Г.М., Крылов В.Ю., Савченко Т.Н. Математические методы в современной психологии: статус, разработка, применение. М.: Изд-во «Институт психологии РАН», 1995.

Классификация и кластер. М.: Мир, 1980.

Многомерный статистический анализ в экономике. М.: Юнити, 1999.

Крылов В.Ю., Острякова Т.В. Математические методы обработки данных в психологических исследованиях: новые методы кластерного анализа на основе психологической теории развития понятий Л. С. Выготского //Психологический журнал. 1995. Т. 16. № 1.
Благуш П. Факторный анализ в обобщении. М.: Финансы и статистика, 1989.
Лазарфельд П. Логические и математические основания латентно-структурного анализа // Математические методы в современной буржуазной социологии. М.: «Прогресс», 1996.
Плюта В. Сравнительный многомерный анализ моделирований. Психологические измерения. М.: Мир., 1967.
Савченко Т.Н., Рассказова Н.П. Применение метода кластерного анализа для изучения отношения школьников к компьютеру // Математические методы в исследованиях индивидуальной и групповой деятельности, М: Институт психологии АН СССР, 1990.

APPLICATION OF METHODS OF CLUSTER ANALYSIS TO DATA PROCESSING IN PSYCHOLOGICAL STUDIES

SAVCHENKO T.N., *Institute of Psychology RAS, Moscow*

The author describes the general provisions of the basic characteristics and goals of cluster analysis, and also conducts a comparative analysis of the most used methods of cluster analysis. The article presents standard methods implemented in the most commonly used statistical packages, their development and improvement and describe original methods that are not realized in statistical packages.

Keywords: multivariate data analysis, classification, cluster, models of latent groups, hierarchical agglomeration method, metrics, dendrogram, strategy of unification, distance matrix, iterative procedure, dendrite, evaluation of bound, associative and chain clusters, cluster-collection, method of latent classes.

Transliteration of the Russian references

Vygotskij L.S. Myshlenie i rech'// Sobr. soch.: V 6 T. T. 2. M.: Pedagogika, 1982.
Glinskij V.V., Ionin V.G. Statisticheskij analiz dannyh, M.: Filin, 1998.
Golovina G.M., Krylov V.Yu., Savchenko T.N. Matematicheskie metody v sovremennoi psihologii: status, razrabotka, primenenie. M.: Izd-vo «Institut psihologii RAN», 1995.
Klassifikatsia i klaster. M.: Mir, 1980.
Mnogomernyi statisticheskij analiz v ekonomike. M.: Yuniti, 1999.
Krylov V.Yu., Ostryakova T.V. Matematicheskie metody obrabotki dannyh v psihologicheskih issledovaniyah: novie metody klasternogo analiza na osnove psihologicheskoi teorii razvitiya ponyatij L.S. Vygotskogo // Psihologicheskij zhurnal. 1995. T. 16. № 1.
Blagush P. Faktornyi analiz v obobschenii. M.: Finansy i statistika, 1989.
Lazarfel'd P. Logicheskie i matematicheskie osnovaniya latentno-strukturnogo analiza // Matematicheskie metody v sovremennoi burzhuaznoi sociologii. M. «Progress», 1996.
Plyuta V. Sravnitel'nyi mnogomernyi analiz modelirovanij. Psihologicheskie izmereniya. M.: Mir., 1967.
Savchenko T.N., Rasskazova N.P. Primenenie metoda klasternogo analiza dlya izucheniya otnosheniya shkol'nikov k komp'yuteru // Matematicheskie metody v issledovaniyah individual'noi i gruppovoi deyatel'nosti, M: Institut psihologii AN SSSR, 1990.