Применение методов кластерного анализа для анализа данных психологических исследований

Т. Н. Савченко

Аннотация: *особенности применения методов кластерного анализа в анализе данных психологических исследований.*

Ключевые слова: *метрика нормированного Евклида, методы кластерного анализа, кластерный анализ, матрицы смешения, матрица расстояний, иерархический агломеративный метод дендритного кластерного анализа, дискриминантный анализ.*

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

Методы кластерного анализа можно расклассифицировать на

– внутренние (признаки классификации равнозначны) и

– внешние (существует один главный признак, остальные определяют его).

Внутренние методы можно разделить на

- иерархические (процедура классификации имеет древовидную структуру);

- неиерархические.

Иерархические классифицируются на

- агломеративные (объединяющие);

- дивизивные (разъединяющие).

Необходимость в использовании методов кластерного анализа возникает, когда задано множество характеристик и множество людей протестировано по ним. Задача в том, чтобы выделить классы людей, близких по всему множеству характеристик (профилю).

На первом этапе необходимо из матрицы смешения (оценки людей по различным характеристикам) построить матрицу расстояний.

Для подсчета матрицы расстояния необходимо выбрать метрику или метод вычисления расстояния между объектами в многомерном пространстве.

Если объект описывается k-признаками, то он может быть представлен как точка в k-мерном пространстве. И чтобы оценить расстояние между объектами в k-мерном пространстве, вводится понятие метрики.

Пусть объекты i и j принадлежат множеству M, 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 $d_{iz} \leqslant dij + dzj$

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, отражает среднее различие между объектами в среднем.

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

Если дисперсия по характеристикам очень сильно отличается, 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 стенах, а другая в IQ, то, чтобы обеспечить одинаковое влияние 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 (Манхеттенская метрика):

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

Название получила в честь района Манхэттен, который образуют улицы, расположенные в виде пересечения параллельных прямых под прямым углом. Манхеттеновская метрика, как правило, применяется для номинальных или качественных переменных.

4. Метрика на основе корреляции:

$d_{ij}=1-|r_{ij}|$;

Расстояния, вычисляемые на основе коэффициента корреляции, отражают согласованность колебаний оценок, в отличие от метрики Евклида, которая определяет схожесть в среднем.

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. Выбор метрики определяется задачей исследования и типом данных. Помимо приведенных выше методов разработаны метрики для ранговых и дихотомических данных и т. д.

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 конце все объекты образуют один класс.

Алгоритм агломеративного метода можно представить в следующем виде:

• На входе метода – матрица смешения или матрица расстояния.

1. На первом шаге находится минимальное расстояние между объектами, и эти объекты объединяются в один класс.

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 У – расстояния, на которых происходит объединение объектов в кластеры. Для определения естественного числа кластеров вводится оценка разбиения на классы, которая определяется отношением средних внутрикластерных расстояний к межкластерным расстояниям (А. Дрынков, Т. Савченко, 1980). Глобальный минимум оценки характеризует естественное число классов, а локальные – под- и над- структуры. 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 «на глаз» по соотношению межкластерных и внутрикластерных расстояний). Далее проводится либо метод k-means, либо дискриминантный анализ, либо авторы самостоятельно используют 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.

Обычно используются следующие стратегии объединения (название несколько не соответствует содержанию, фактически используя выбранные формулы, мы пересчитываем расстояния от объектов до вновь образованного класса):

1. стратегия «ближайшего соседа», она 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, соответствует расстоянию 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 не изменяет пространство. 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 близости границ, а вторая – по дальним границам.

Таким образом, обычно в стандартных ситуациях используется стратегия «группового среднего». Если исследуемая группа достаточна разнородная,

Таблица 1

Матрица смешения для коллектива из 9 человек

№	d (зав. от гр. станд.)	Resp (ответст.)	Lab (труд. акт.)	Wtg (работоспособн.)	Goal (понимание цели)	Mot (мотив.)
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

т.е. люди, входящие в нее, значимо отличаются друг от друга по множеству характеристик, но все же необходимо выделить среди них группы более сходные по всему профилю характеристик, то используется стратегия «дальнего соседа» (сужающая пространство). Если же группа достаточно гомогенная, тогда, чтобы выделить среди очень схожих людей подгруппы, следует использовать стратегию «дальнего соседа».

Рассмотрим фрагмент результатов исследования успешности деятельности команды – малой группы, ориентированной на выполнение деловой задачи, 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 в таб. 1.

Используя метрику Евклида, получим симметричную матрицу расстояний. Эта матрица является входной в методы кластерного анализа (см. таб. 2).

Результат применения агломеративного иерерхического метода КА к полученной матрице при использовании пакета STATISTICA – дерево классификации представлен на Рис.1. По горизонтальной оси представленного дерева откладываются на одинаковом расстоянии номера объектов (членов команды), по вертикальной оси – расстояния, 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

Таблица 2

Матрица расстояний, полученная с использованием метрики Евклида
Euclidean distances (dip12.sta)

	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

более эффективным при использовании оценок разбиения на классы, однако в стандартных пакетах (как отмечалось выше) такая оценка, к сожалению, не предусмотрена. Для получения большей информации о полученных классах используются другие методы кластеризации: 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.

В социальной психологии при исследовании взаимоотношений в коллективах помимо разбиения на классы необходимо также проанализировать, через какие объекты классы связаны друг с другом. На эти вопросы можно ответить с помощью **дендритного кластерно-**

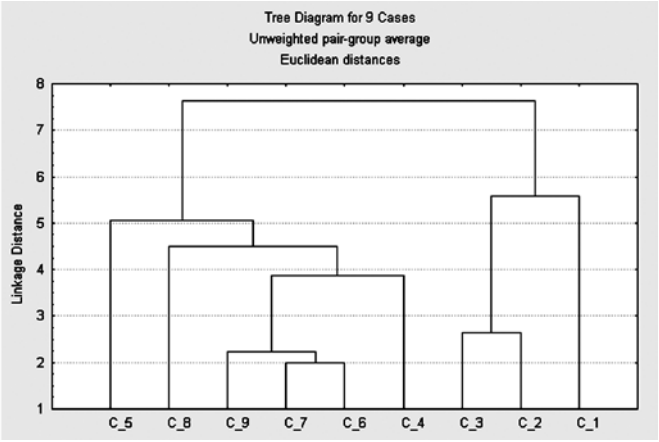

Рис.1. Дерево классификации.

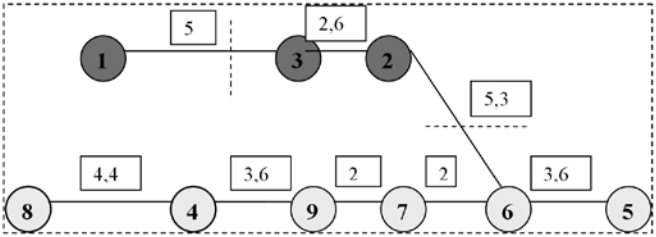

Рис. 2. Дендрит (форма простого дерева).

го анализа, который часто применяется совместно с иерархическим (Плюта, 1989).

Главная роль в нем принадлежит дендриту. Дендрит – 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. 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 есть один и тот же элемент, то пары объединяются через этот элемент. Например, две пары

i_____j

j_____k

объединяются в связку

i_____j_____r

На этом заканчивается построение скоплений (плеяд) первого порядка. Затем находятся минимальные расстояния между объектами скоплений первого порядка, и эти скопления объединяются через эти объекты 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 берется из первого метода, а взаимосвязи между ними анализируются с помощью дендрита.

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

Для рассмотренного ранее случая Ср =4.8. Это означает, что выделяется три класса, что несколько отличается от результата агломеративного метода. Из первого класса, в который входили объекты 1, 3, 2, отделился первый сотрудник. Во второй вошли – 8, 4, 9, 7, 6, 5 (аналогично агломеративному методу). Над дугами дендрита указаны расстояния между объектами.

Применение данного метода позволяет получить дополнительную информацию о том, через какие объекты классы взаимосвязаны друг с другом. В данном случае это 2 и 6 объекты (члены коллектива). Данная структура аналогична социометрии, однако получена на

основе результатов тестирования. Анализ дендрита позволит выделить группы совместимых людей, которые наиболее эффективно смогут совместно решить поставленные задачи, либо выделить людей, которые лучше работают в одиночку, например, объект 1. 8-й объект также находится на границе отделимости, поэтому возможно ему также лучше давать индивидуальные задания.

Помимо агломеративных иерархических методов существует также большое количество **итеративных методов кластерного анализа**. Основное отличие их 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 др. Эти методы требуют от исследователя интуиции, творческого подхода. Необходимо представлять еще до проведения классификации, какое количество классов должно быть образовано, когда закончить процесс классификации и т. д. От верно выбранных начальных условий будет зависеть результат классификации. Неверно выбранные условия могут приводить к размытым классам. Поэтому эти методы используются, если есть теоретическое обоснование, например, количества ожидаемых классов, или после проведения иерархических методов классификации, которые позволят выработать наиболее оптимальную стратегию поведения.

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 случайным образом или исходя 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 этого используется одна из метрик. К сожалению, в основных статистических пакетах используется только метрика Евклида. Рассматриваемый объект относится к такому классу, к эталону которого он ближе всего. Если есть два одинаковых минимальных расстояния, то объект присоединяется к классу с минимальным номером.

Эталон, к которому присоединен новый объект, пересчитывается, а его вес возрастает на единицу.

Пусть эталоны представлены таким образом:

$e(1,O)=(x_{11}, x_{12}, ...,x_{1i}, ..., x_{1m})$

.....

$e(k,O)=(x_{k1}, x_{k2}, ...,x_{ki}, ..., x_{km})$

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

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

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

Остальные эталоны остаются неизменными.

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

Чтобы получить устойчивое разбиение, полученные новые эталоны после отнесения всех объектов принимаются за начальные эталоны, и процедура повторяется с самого начала. Веса классов продолжают накапливаться. Новое распределение по классам сравнивается с предыдущим, если различие не превышает заданного уровня, т. е. распределения можно считать неизменившимися, 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. Независимо от модификации происходит минимизация внутри кластерной дисперсии. Это относится к большинству итерационных методов кластерного анализа.

Обычно в методе k-средних реализована процедура построения усредненных профилей каждого класса. Это 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. Это позволяет проверить значимость различия классов по исследуемым характеристикам.

Анализ профилей показывает, что в первый класс попали сотрудники, у которых слабая зависимость от группы, средний уровень ответственности и высокие трудовая активность, работоспособность, понимание цели. Во вторую группу (более многочисленную) вошли сотрудники, у которых есть зависимость от групповых стандартов, низкая ответственность и достаточно невысокие трудовая активность, работоспособность и понимание общей цели. На людей, которые отнесены к первой группе, можно возлагать ответственность, они могут самостоятельно принимать решения и

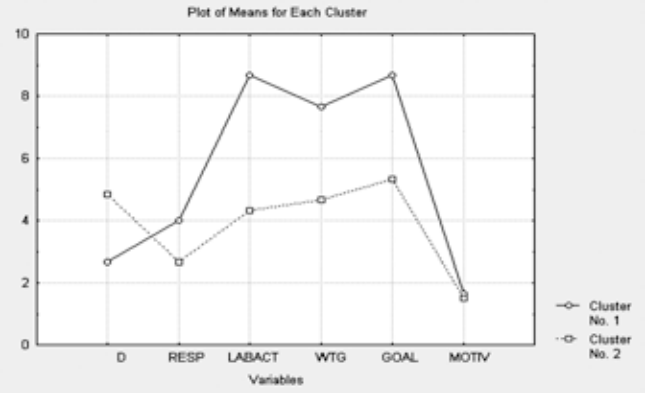

Рис. 3. Усредненные профили классов.

Таблица 4

Номера объектов и расстояния от центра класса, соответственно 1-го и второго классов

Members of Cluster Number 1 (dip12.sta)			
	1.	**3.**	**2.**
Distance	1.484488	1.097134	0.693889

Members of Cluster Number 2 (dip12.sta)						
	8	**4**	**9**	**7**	**6**	**5**
Distance	1.357421	1.566430	0.535758	0.855267	1.272938	0.822147

Таблица 5

Анализ отделимости классов

Analysis of Variance (dip12.sta)					
	Межкл. Расст.	**Внутрикл. Расст.**	**Ст.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

т.д., вторая группа – это исполнители, и при этом необходим постоянный контроль за выполнением порученных заданий. Можно заметить, что мотивация низкая у обеих групп, что связано, возможно, с невысокой оплатой труда. Данные объяснения мы приводим с целью объяснения возможностей метода.

В таб. 5 представлены результаты сравнительного анализа классов, которые показывают, что классы значимо отличаются друг от друга по трем характеристикам: трудовая активность, работоспособность и понимание цели.

В таблице выделены жирным шрифтом те характеристики, по которым наблюдается значимое различие между классами.

Дискриминантный анализ

Рассмотрим возможности данного метода применительно к задаче исследования структуры команды (см. таб. 6).

Результаты иерархической классификации показали, что из исходных данных выделяется два класса. Необходимо отнести новый объект (10) к одному из классов.

1. Необходимо найти среднее значение всех характеристик для объектов каждого класса (найти вектор средних значений для каждого класса) (таб. 7).

Таблица 6

Исходные данные

№	d	resp	lab	wtg	goal	mot
1	2	7	9	8	10	3
2	4	2	8	8	8	1
3	2	3	9	7	8	1
4	7	3	5	6	4	0
5	2	2	5	3	7	2
6	4	3	5	5	5	2
7	5	4	4	5	5	3
8	6	1	4	4	7	0
9	5	3	3	5	4	2
Новый объект						
10	1	8	8	7	9	5

2. Находится ковариационная матрица для объектов первого и второго класса.

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

На примере первой ковариации первого класса:

$$Cov_{12} = \frac{(2-2{,}6)(7-4)+(4-2{,}6)(2-4)+(2-2{,}6)(3-4)}{3-1} = -4$$

Таблица 7

	1 КЛАСС	2 КЛАСС
1	2,6	4,8
2	4	2,6
3	8,67	4,3
4	7,6	4,6
5	8,6	5,3
6	1,6	1,5

В результате ковариационного анализа двух классов были получены две матрицы:

Для первого класса:

	1	2	3	4	5	6
1	1,33	-4	-1,33	0,67	-1,33	-1,33
2	-4	7	2	1	6	6
3	-1,33	2	0,33	-0,33	0,44	0,67
4	0,67	1	-0,33	0,33	0,67	0,67
5	-1,33	6	0,44	0,67	2,68	2,67
6	-1,33	6	0,67	0,67	2,67	2,68

Для второго класса:

	1	2	3	4	5	6
1	14,84	0,69	-1,69	6,69	-5,57	-6,52
2	0,69	5,36	-0,34	3,36	-5,35	4,01
3	-1,69	-0,34	3.34	-0,35	1,35	-1
4	6,69	3,36	-0,35	7,76	-6,32	-2,1
5	-5,57	-5,35	1,35	-6,32	9,36	-1
6	-6,52	4,01	-1	-2,1	-1	7,5

Подсчитывается дисперсия для каждого класса (значения находятся в окошках, выделенных серым цветом, см. выше).

3. Оценивается суммарная внутривыборочная дисперсия (S*).

$$S^* = \frac{1}{3+6-2} \times$$

	1	2	3	4	5	6
1	16,17	-3,31	-3,02	7,36	-6,9	-7,85
2	-3,31	12,36	1,66	4,36	0,65	10,01
3	-3,02	1,66	3,67	-0,68	1,79	-0,33
4	7,36	4,36	-0,68	8,09	-5,65	-1,45
5	-6,9	0,65	1,79	-5,65	12,04	1,67
6	-7,85	10,01	-0,33	-1,45	1,67	10,18

=

	1	2	3	4	5	6
1	2,26	0,46	-0,42	1,03	-0,96	-1,09
2	-0,46	1,73	0,23	0,61	0,09	1,4
3	-0,42	0,23	0,51	-0,09	0,25	-0,05
4	1,03	0,61	-0,09	1,13	-0,79	-0,2
5	-0,96	0,09	0,25	-0,79	1,69	0,23
6	-1,09	1,4	-0,05	-0,2	0,23	1,43

4. Определяется обратная суммарная матрица внутривыборочной дисперсии.

S_*^{-1}=

	1	2	3	4	5	6
1	1,055	2,222	-0,536	-2,899	-0,563	-1,705
2	0,462	-0,84	0,848	0,486	0,241	1,233
3	0,515	2,166	1,232	-2,523	-0,928	1,889
4	-1,194	-0,758	-0,502	2,906	0,782	0,095
5	-0,089	0,3	-0,608	0,39	0,861	-0,467
6	0,217	2,437	-1,168	-2,431	-0,727	-1,785

5. Определяются дискриминативные множители.

$A = S_*^{-1}(\overline{x}_1 - \overline{x}_2)$

$\overline{x}_1 - \overline{x}_2 =$

-2/2
1,4
4,37
3
3,3
0,1

$A =$

-3,2
2,36
4,11
5,75
2,64
-4,16

6. Находятся средние значения дискриминантной организации для объектов первого и второго классов. (Находятся значения дискриминанты для каждого объекта, после чего подсчитываются средние значения.)

На примере первого класса:

$f_{11}=2(-3,26)+7\times2,36+9\times4,11+8\times5,75+10\times2,64+3(-4,16)=106,91$

$f_{12}=-13,04+4,72+32,88+46+21,12-4,16=87,52$

$f_{13}=-6,52+7,08+36,99+40,25+21,12-4,16=94,76$

$$\overline{f_1} = \frac{f_{11} + f_{12} + f_{13}}{3}$$

$\overline{f}_1 = 96{,}39$
$\overline{f}_2 = 43{,}02$

7. В простейшем случае граница между классами находится как полусумма дискриминантной функции.

$$C = \frac{\overline{f_1} + \overline{f_2}}{2} = \frac{139{,}41}{2} = 69{,}70$$

8. Находим функцию объекта.

fоб=–3,26+18,88+32,88+40,25+23,76–20,8=91,71

Так как Fоб > C (91,71> 69,70) следовательно, объект принадлежит к первому классу.

Таким образом, с помощью дискриминантного анализа можно доказать принадлежность объекта к данному классу, подсчитав вероятности принадлежности объекта к классам и выбрав максимальную. Т. е. проверить верность проведенной классификации и подправить ее при необходимости. Данный метод можно использовать с классами, выделенными на основе теоретических посылок, или при свободной классификации, а также совместно с агломеративным иерархическим методом с целью доказательства правильности классификации «на глаз». Другая цель применения данного метода – это прогноз о принадлежности нового объекта к одному из существующих классов. Метод дискриминантного анализа чрезвычайно полезен, например, в ситуации профотбора. Получив профили успешных представителей той или иной профессиональной сферы и, выделив профиль данного претендента, можно спрогнозировать успешность его деятельности по близости его дискриминантной функции к функции соответствующего класса. В данной статье мы попытались описать наиболее используемые в психологии методы кластерного анализа, проанализировав их достоинства и недостатки, и соотнести их друг с другом. На одном примере мы продемонстрировали совместное применение различных методов. Все эти методы реализованы в статистических пакетах (STATISTIKA, SPSS), исключение составляет дендритный метод, однако применение стратегии ближайшего соседа позволит в определенной степени реализовать данный метод с использованием алгоритма агломеративного метода. Кластерный анализ эффективно применять совместно с методами факторного анализа и методов многомерного шкалирования – но это тема следующей статьи.

Литература

1. Глинский В. В., Ионин В. Г. Статистический анализ данных, М.: Филин, 1998
2. Головина Г. М., Крылов В. Ю., Савченко Т. Н. Математические методы в современной психологии: статус, разработка, применение. М.: Изд-во Институт психологии РАН, 1995.
3. Ермолаев О. Ю. Математическая статистика для психологов, М.: Московский психолого-социальный институт изд. «Флинта», 2002
4. Классификация и кластер. М.: Мир, 1980.
5. Многомерный статистический анализ в экономике. М.: Юнити, 1999
6. Плюта В. Сравнительный многомерный анализ моделирований. Психологические измерения. М.: Мир., 1967.
7. Синергетический подход к моделированию психологических систем. / Под ред. Т. Н. Савченко. М.: Изд-во Институт психологии РАН, 1998.
8. Суходольский Г. В. Основы математической статистики для психологов. СПб., 1997.
9. Наследов А. Д. Математические методы психологического исследования. СПБ., Речь, 2006.