ДИСЦИПЛИНА	**Интеллектуальные системы и технологии**
	(полное наименование дисциплины без сокращений)
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**корпоративных информационных систем**
	полное наименование кафедры
ВИД УЧЕБНОГО МАТЕРИАЛА	**Лекция**
	(в соответствии с пп.1-11)
ПРЕПОДАВАТЕЛЬ	**Демидова Лилия Анатольевна**
	(фамилия, имя, отчество)
СЕМЕСТР	**1 семестр (осенний), 2024 – 2025 учебный год**
	(семестр обучения, учебный год)

ЛЕКЦИЯ 7
АЛГОРИТМЫ КЛАСТЕРНОГО АНАЛИЗА

Необходимость анализа больших объемов объективной и субъективной информации, связанных с неформализуемыми и плохо формализуемыми задачами различной физической природы, требует развития новых научных направлений, в том числе прикладной статистики и методов анализа данных.

Применение методов прикладной статистики основано на предположении о вероятностной интерпретации анализируемой информации и получении с помощью этих методов закономерностей, имеющих стохастический характер.

Методы анализа данных, в том числе кластерный анализ, не используют априорных предположений о вероятностной природе исходной информации и применяют только эвристические соображения о характере и особенностях исследуемой совокупности объектов. В основе этой теории лежит нечетко-возможностная интерпретация неопределенности.

Кластерный анализ – это совокупность методов, алгоритмов, подходов и процедур, разработанных для решения проблемы формирования однородных классов в произвольной проблемной области. Кластерный анализ занимает одно из центральных мест среди методов анализа данных и представляет собой совокупность подходов, методов, алгоритмов, предназначенных для нахождения некоторого разбиения исследуемой совокупности объектов на подмножества относительно сходных, похожих между собой объектов. Исходным допущением для выделения таких подмножеств – кластеров – служит неформальное предположение о том, что объекты, относимые к одному кластеру, должны иметь большее сходство между собой, чем с объектами из других кластеров.

При выявлении во множестве данных кластеров следует соблюдать следующие условия:

– каждый кластер должен представлять собой концептуально однородную категорию и содержать похожие объекты с близкими значениями свойств

или признаков;

– совокупность всех кластеров должна быть исчерпывающей (охватывать все объекты исследуемой совокупности);

– кластеры должны быть взаимно исключающими (ни один объект исследуемой области не должен одновременно принадлежать двум различным кластерам).

Под задачей кластерного анализа заданного множества объектов понимается задача нахождения некоторого теоретико-множественного разбиения этого исходного множества объектов на непересекающиеся подмножества – кластеры – таким образом, чтобы элементы, относимые к одному подмножеству, отличались друг от друга в значительно меньшей степени, чем элементы из разных подмножеств.

Методы и алгоритмы кластерного анализа используются при поиске закономерностей в больших наборах многомерных данных, таких как хранилища данных. При этом проблема кластер-анализа приобретает самостоятельное значение в контексте интеллектуального анализа данных (Data Mining).

Кластеризация – это:

– группировка объектов по похожести их свойств; каждый кластер состоит из схожих объектов, а объекты разных кластеров существенно отличаются;

– процедура, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$.

Постановка задачи кластеризации сложна и неоднозначна, так как:

– число кластеров в общем случае неизвестно;

– выбор меры «похожести» или близости свойств объектов между собой, как и критерия качества кластеризации, всегда носит субъективный характер.

Алгоритмы кластеризации разбивают заданное множество объектов на группы (кластеры), в одном кластере размещая близкие, а в разных далекие по своим характеристикам (признакам) объекты.

Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

Кластерный анализ – группа методов, используемых для классификации объектов или событий в относительно гомогенные (однородные) группы, которые называют кластерами (clusters).

В **факторном анализе** группируются столбцы, т.е. цель – анализ структуры множества признаков и выявление обобщенных факторов.

В **кластерном анализе** – группируются строки, т.е. цель – анализ структуры множества объектов.

Кластерный анализ выполняет **классификацию объектов.**

Каждый объект – точка в пространстве признаков.

Задача кластерного анализа – выделение «сгущений» точек, разбиение совокупности на однородные подмножества объектов (сегментация).

Цели кластеризации

1. Упростить дальнейшую обработку данных, разбив множество объектов на группы схожих объектов, чтобы работать с каждой группой в отдельности (задачи классификации, регрессии, прогнозирования).
2. Сократить объем хранимых данных, оставив по одному представителю от каждого кластера (задачи сжатия данных).
3. Выделить нетипичные объекты, которые не подходят ни к одному из кластеров (задачи одноклассовой классификации).
4. Построить иерархию множества объектов (задачи таксономии).

Задачи, для решения которых используется кластеризация

1. *Изучение данных.* Разбиение множества объектов на схожие группы помогает выявить структуру данных, увеличить наглядность их представления, выдвинуть новые гипотезы, понять, насколько информативны свойства объектов.

2. *Облегчение анализа.* При помощи кластеризации можно упростить дальнейшую обработку данных и построение моделей: каждый кластер обрабатывается индивидуально и модель создается для каждого кластера в отдельности. В этом смысле кластеризация является подготовительным этапом перед решением других задач Data Mining: классификации, регрессии, ассоциации, последовательных шаблонов.

3. *Сжатие данных.* В случае, когда данные имеют большой объем (сотни тысяч и миллионы строк), кластеризация позволяет сократить объем хранимых данных, оставив по одному наиболее типичному представителю от каждого кластера.

4. *Обнаружение аномалий.* Кластеризация применяется для выделения нетипичных объектов, которые не присоединяются ни к одному из кластеров. Эту задачу также называют обнаружением аномалий (outlier detection). Интерес здесь представляют кластеры (группы), в которые попадает крайне мало, например, 1 – 3, объекта.

5. *Прогнозирование.* Кластеры используются не только для краткого описания имеющихся объектов, но и для распознавания новых. Каждый новый объект относится к тому кластеру, присоединение к которому наилучшим образом

удовлетворяет критерию качества кластеризации. Значит, можно прогнозировать поведение объекта, предположив, что оно будет схожим с поведением других объектов кластера.

Представление данных в двумерном случае в кластерном анализе в теории (в идеальном случае) и на практике показано на рисунках 1 и 2 соответственно.

Рисунок 1. Представление данных в теории (в идеальном случае)

Рисунок 2. Представление данных на практике

Каждый алгоритм кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов: например, гиперсферические, цепочечные и т.п.

Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Выявление кластеров может быть реализовано с применением иерархических алгоритмов кластерного анализа и итеративных алгоритмов кластерного анализа. Обычно иерархическую кластеризацию реализуют при малом числе объектов, а итеративную – при большом числе объектов.

Процесс кластерного анализа

В общем виде этапы процесса кластерного анализа представлены на рисунке 3.

Рисунок 3. Этапы кластерного анализа

Постановка задачи кластеризации

Дано:

набор тестовых примеров (объектов) $X=\{x_1,...,x_n\}$ и функция расстояния между ними.

Требуется: разбить набор X на непересекающиеся подмножества (кластеры) так, чтобы каждое подмножество состояло из похожих объектов, а объекты разных подмножеств существенно различались.

Иерархические алгоритмы кластерного анализа

Алгоритмы иерархической кластеризации, называемые также алгоритмами таксономии, – это алгоритмы упорядочивания данных, реализующие создание иерархии (дерева) вложенных кластеров.

Выделяют два типа алгоритмов иерархической кластеризации: дивизимные (нисходящие) алгоритмы (divisive algorithms) создают новые кластеры посредством деления более крупных кластеров на более мелкие (при этом дерево иерархии формируется от ствола к листьям); агломеративные (восходящие) алгоритмы (agglomerative algorithms) создают новые кластеры посредством объединения более мелких кластеров (при этом дерево иерархии формируется от листьев к стволу). Наибольшее применение в решении различных прикладных задач находят агломеративные алгоритмы кластеризации.

Меры сходства. Меры объединения (связи) кластеров

При реализации агломеративного алгоритма кластеризации сначала отдельным кластером считается каждый объект. При этом для любых двух одноэлементных кластеров U и V, состоящих соответственно из объектов x_i и x_j ($i = \overline{1,n}$; $j = \overline{1,n}$; n – число объектов), расстояние между кластерами вычисляется с помощью метрики расстояния:

$$D(U,V) = d(x_i, x_j), \qquad (1)$$

где в качестве метрики $d(x_i, x_j)$ может быть выбрана та или иная метрика вычисления расстояния между объектами.

Так, например, для вычисления расстояния между объектами $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ и $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,p})$ в p-мерном пространстве могут быть использованы такие метрики расстояний (рассматриваемые как меры сходства), как:

– метрика евклидова расстояния:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{i,r} - x_{j,r})^2}\ ; \qquad (2)$$

– метрика квадрата евклидова расстояния:

$$d(x_i, x_j) = \sum_{r=1}^{p} (x_{i,r} - x_{j,r})^2\ ; \qquad (3)$$

– метрика манхэттенского расстояния:

$$d(x_i, x_j) = \sum_{r=1}^{p} | x_{i,r} - x_{j,r} |; \qquad (4)$$

– метрика расстояния Чебышева:

$$d(x_i, x_j) = \max_{r} | x_{i,r} - x_{j,r} |; \qquad (5)$$

– метрика расстояния Минковского:

$$d(x_i, x_j) = (\sum_{r=1}^{p} | x_{i,r} - x_{j,r} |^h)^{\frac{1}{h}} \ (h \geq 1); \qquad (6)$$

– косинусная метрика расстояния:

$$d(x_i, x_j) = \frac{\sum_{r=1}^{p} (x_{i,r}, x_{j,r})}{\sqrt{(\sum_{r=1}^{p} x_{i,r}^2) \cdot (\sum_{r=1}^{p} x_{j,r}^2)}}\ . \qquad (7)$$

Итерационный процесс слияния кластеров осуществляется следующим образом: на каждой итерации на основе двух самых близких кластеров U и V формируется новый кластер $W = U \cup V$. При этом расстояние от нового кластера W до любого другого кластера S вычисляется на основе уже известных расстояний $D(U,V)$, $D(U,S)$ и $D(V,S)$ с помощью универсальной формулы Ланса и Уильямса:

$$D(U \cup V, S) =$$

$$= \alpha_U \cdot D(U,S) + \alpha_V \cdot D(V,S) + \beta \cdot D(U,V) + \gamma \cdot | D(U,S) - D(V,S) |, \qquad (8)$$

где α_U, α_V, β, γ – некоторые числовые параметры.

Формула (2) описывает практически все возможные методы вычисления расстояний между кластерами при определённых комбинациях значений параметров α_U, α_V, β, γ (таблица 1, где $|U|$, $|V|$, $|W|$ и $|S|$ – мощности кластеров U, V, W и S соответственно).

Так, например,

– расчет расстояния между кластерами по методу одиночной связи реализуется как:

$$D_{single}(W,S) = \min_{w \in W, s \in S} d(w,s), \quad (9)$$

– расчет расстояния между кластерами по методу полной связи как:

$$D_{complete}(W,S) = \max_{w \in W, s \in S} d(w,s), \quad (10)$$

– расчет по методу средней связи как:

$$D_{average}(W,S) = \frac{1}{|W| \cdot |S|} \cdot \sum_{w \in W} \sum_{s \in S} d(w,s), \quad (11)$$

– расчет расстояния по методу центроида как:

$$D_{centroid}(W,S) = d^2\left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right), \quad (12)$$

– расчет расстояния Уорда как:

$$D_{Ward}(W,S) = \frac{|W| \cdot |S|}{|W| + |S|} \cdot d^2\left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right). \quad (13)$$

Следует отметить, что при выборе метода слияния кластеров необходимо учитывать его совместимость с метрикой для вычисления расстояния между объектами. Так, метод центроида, метод медианы и метод Уорда предполагают использование метрики евклидова расстояния.

В результате применения того или иного метода слияния кластеров формируется матрица сходства (или различия) кластеров, которая определяет уровень сходства (различия) между парами кластеров и используется для анализа и визуализации результатов кластеризации посредством построения дендрограммы.

Как и большинство визуальных способов представления зависимостей дендрограммы теряют наглядность при значительном увеличении числа кластеров.

Алгоритм иерархической кластеризации может быть описан следующей последовательностью шагов.

Таблица 1. Значения параметров α_U, α_V, β, γ в универсальной формуле (8) для определения метода слияния кластеров

Метод слияния кластеров	α_U	α_U	β	γ
Метод одиночной связи (метод ближайшего соседа, (single-linkage clustering method)	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Метод полной связи (метод дальнего соседа, (complete-linkage clustering method)	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Метод средней связи (метод группового среднего расстояния, average-linkage clustering method)	$\frac{\lvert U\rvert}{\lvert W\rvert}$	$\frac{\lvert V\rvert}{\lvert W\rvert}$	0	0
Метод взвешенной средней связи (метод взвешенного группового среднего расстояния, weighted average-linkage clustering method)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Метод центроида (centroid-linkage clustering method)	$\frac{\lvert U\rvert}{\lvert W\rvert}$	$\frac{\lvert V\rvert}{\lvert W\rvert}$	$-\alpha_U \cdot \alpha_V$	0
Метод медианы (median-linkage clustering method)	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Метод расчета расстояния Уорда (Ward-linkage clustering method)	$\frac{\lvert U\rvert+\lvert S\rvert}{\lvert W\rvert+\lvert S\rvert}$	$\frac{\lvert V\rvert+\lvert S\rvert}{\lvert W\rvert+\lvert S\rvert}$	$\frac{-\lvert S\rvert}{\lvert W\rvert+\lvert S\rvert}$	0

Шаг 1. Принять номер шага слияния l, равным 1. Выполнить инициализацию множества кластеров одноэлементными кластерами, состоящими из объектов кластеризации:

$C^l = \{\{x_1\},\{x_2\},...,\{x_n\}\}$.

Шаг 2. Выполнить расчет матрицы расстояний между одноэлементными кластерами на основе формулы (1).

Шаг 3. Увеличить номер шага слияния l на 1. Найти в множестве C^{l-1} два ближайших кластера U и V, а затем объединить их в один кластер: $W = U \cup V$. Удалить из множества кластеров C^{l-1} кластеры U и V. Добавить в множество кластеров C^{l-1} новый кластер W: $C^l = (C^{l-1} \setminus \{U,V\}) \cup \{W\}$.

Шаг 3. Для всех кластеров $S \in C^l$ вычислить расстояние $D(W,S)$ по формуле Ланса-Уильямса (2).

Шаг 4. Если число кластеров в множестве C^l больше 1, перейти к шагу 3, в противном случае завершить работу алгоритма.

Пример иерархического кластерного анализа

Рассмотрим возможности иерархического кластерного анализа на примере набора данных «Ирисы Фишера» (iris dataset).

Ирисы Фишера – это набор данных для задачи классификации, на примере которого Рональд Фишер в 1936 году продемонстрировал работу разработанного им метода дискриминантного анализа.

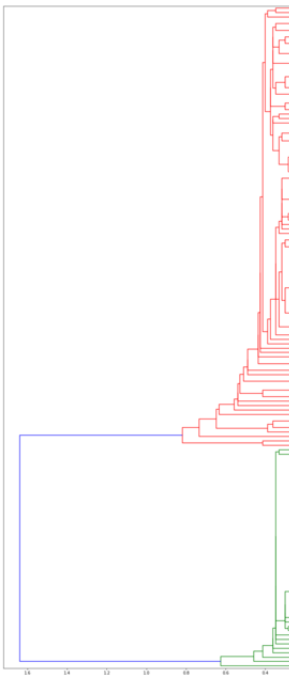
Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов – Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*). Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):

1. Длина наружной доли околоцветника (*sepal length*);
2. Ширина наружной доли околоцветника (*sepal width*);
3. Длина внутренней доли околоцветника (*petal length*);
4. Ширина внутренней доли околоцветника (*petal width*).

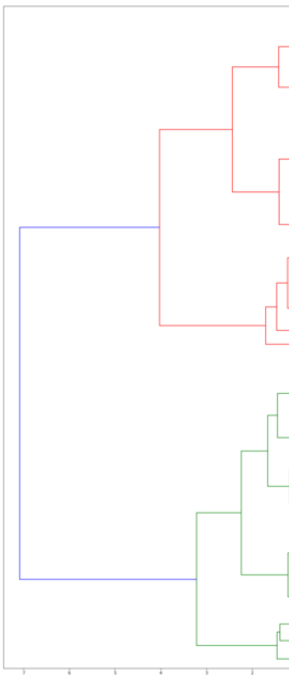
На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений. Это задача многоклассовой классификации, так как имеется три класса – три вида ириса.

Один из классов (Iris setosa) линейно-отделим от двух остальных.

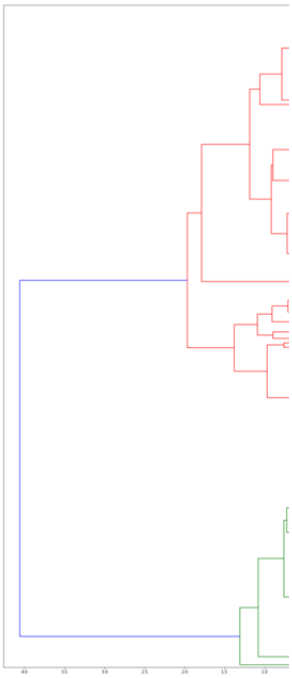
Лучше всего с задачей справился алгоритм с использованием расстояния Уорда. Он точно выделил класс Iris setosa и заметно отделил вид Iris virginica от Iris versicolor.



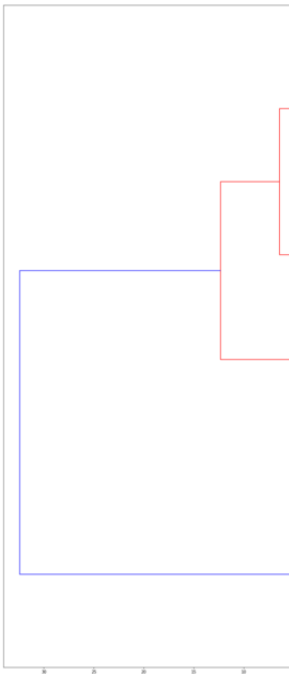
Метод одиночной связи



Метод полной связи



Метод средней связи



Метод Уорда

Рисунок 4. Дендрограммы кластеризации ирисов Фишера в зависимости от функции расстояния между кластерами

Код (Python)

```
# Подключение библиотек
from scipy.cluster.hierarchy import linkage, dendrogram
from sklearn import datasets
import matplotlib.pyplot as plt

# Создание полотна для рисования
fig = plt.figure(figsize=(15, 30))
fig.patch.set_facecolor('white')

# Загрузка набора данных "Ирисы Фишера"
iris = datasets.load_iris()

# Реализация иерархической кластеризации при помощи функции linkage
mergings = linkage(iris.data, method='ward') # метод Уорда

# Построение дендрограммы. Разными цветами выделены автоматически определен-
ные кластеры
R = dendrogram(mergings, labels=[iris.target_names[i] for i in iris.target],
orientation = 'left', leaf_font_size = 12)

# Отображение дендрограммы
plt.show()
```

На рисунке 5 показаны принципы группирования кластеров при выполнении иерархической кластеризации с использованием агломеративных и дивизимых методов.

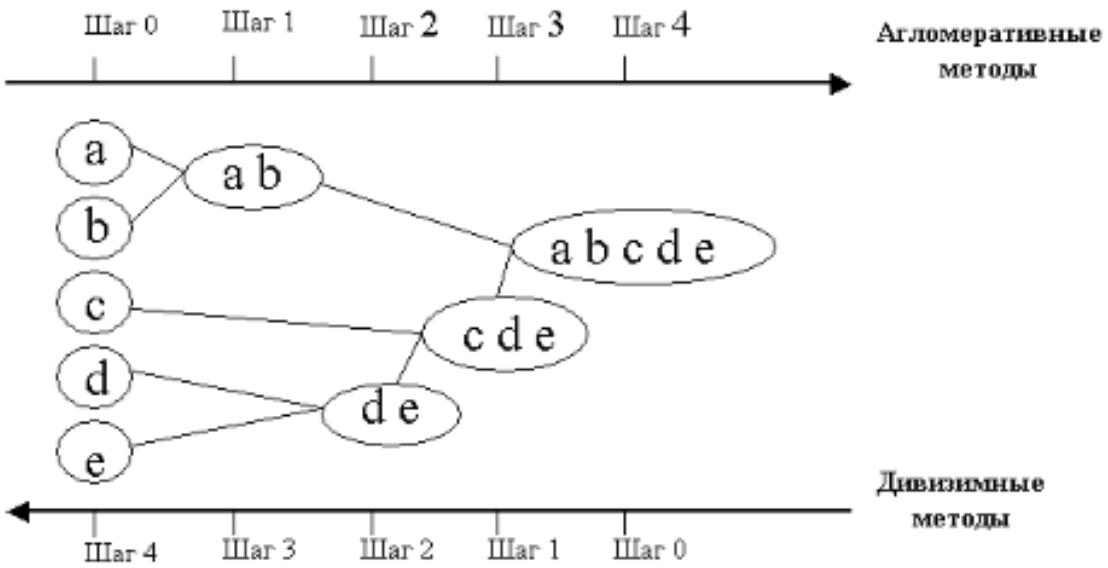


Рисунок 5. Направления иерархической кластеризации при реализации различных методов

Алгоритм k-средних (k-means)

Алгоритм k-средних предполагает минимизацию функции:

$$\sum_{i=1}^{n}\|x_i-\mu_a\|^2\to\min_{\{a_i\},\{\mu_a\}};\ \|x_i-\mu_a\|^2=\sum_{j=1}^{l}\left(f_j(x_i)-\mu_{a_j}\right)^2,$$

где $(f_1(x_i),...,f_l(x_i))$ – вектор признаков объекта x_i.

Алгоритм k-средних может быть реализован в виде алгоритма Ллойда.

Алгоритм Ллойда

Вход: X^n, $k=|Y|$. Выход: центры кластеров μ_a, $a\in Y$.

1. μ_a:= начальное приближение центроидов, для всех $a\in Y$;
2. **повторять**
3. отнести каждый x_i к ближайшему центроиду:

$$a_i:=\arg\min_{a\in Y}\|x_i-\mu_a\|,\ i=\overline{1,n};$$

4. вычислить новые положения центроидов:

$$\mu_a:=\frac{\sum_{i=1}^{n}[a_i=a]x_i}{\sum_{i=1}^{n}[a_i=a]},\ a\in Y;$$

5. **пока** a_i не перестанут изменяться;

Пример кластеризации на основе алгоритма k-means для «Ирисов Фишера»

Результаты кластеризации алгоритмом k-средних для ирисов Фишера и реальные виды ирисов (в случае анализа 3 характеристик) показаны на рисунке 6. Центроиды кластеров отмечены с помощью крупных, полупрозрачных маркеров.

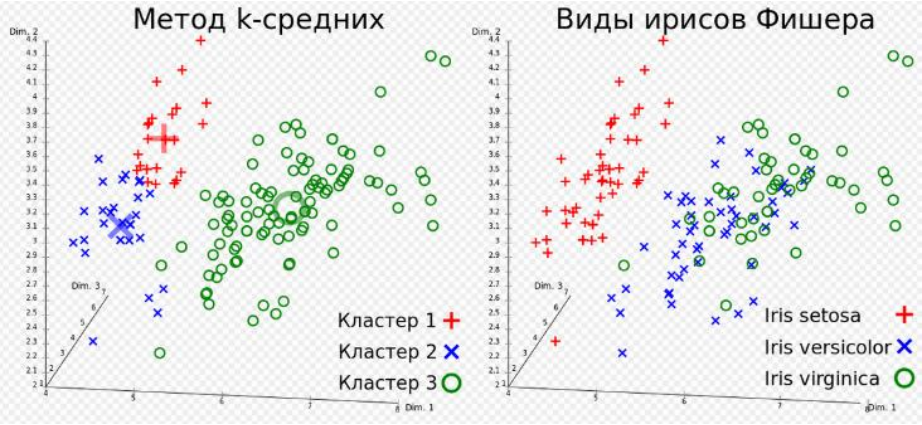


Рисунок 6

FCM-алгоритм (fuzzy c-means)

FCM-алгоритм предполагает минимизацию целевой функции, представляющей собой сумму взвешенных квадратов расстояний:

$$J(U,V)=\sum_{q=1}^{c}\sum_{i=1}^{n}\left(u_q(x_i)\right)^m\cdot d^2\left(v_q,x_i\right) \qquad (14)$$

при

$$\sum_{q=1}^{c}u_q(x_i)=1\ (c\in N\ \text{и}\ c>1;\ i=\overline{1,n}), \qquad (15)$$

где $U=[u_q(x_i)]$ – нечеткое c-разбиение множества объектов на основе функций принадлежности $u_q(x_i)$, определяющих степень принадлежности i-го объекта q-му кластеру; $V=(v_1,...,v_c)$ – центры кластеров; $d(v_q,x_i)$ – расстояние между центром кластера v_q и объектом x_i; m – фаззификатор ($m\in R$, $m>1$); c – число кластеров; n – число объектов; $i=\overline{1,n}$; $q=\overline{1,c}$.

Следует отметить, что условие (15) определяет нечеткое разбиение множества объектов R в виде: $\cup_q R_q=R$ ($R_q\in R$).

Функции принадлежности играют роль весовых коэффициентов, определяя степень вклада объекта в оценку центров кластеров и, соответственно, степень принадлежности i-го объекта q-му кластеру. Размер вклада зависит от выбора фаззификатора m, управляющего степенью «нечеткости».

При малых значениях m нечеткое разбиение вырождается в четкое, при $m \to \infty$ степени принадлежности объектов каждому кластеру становятся равны $1/c$, то есть каждый объект равновероятно принадлежит каждому кластеру. Обычно в качестве значения фаззификатора m выбирается значение, равное 2.

FCM-алгоритм предполагает выполнение следующих шагов.

1. Инициализация начального нечеткого разбиения $U = [u_q(x_i)]$, удовлетворяющего условию (15).

2. Вычисление координат центров кластеров:

$$v_q^l = \frac{\sum_{i=1}^{n} u_q(x_i)^m \cdot x_i^l}{\sum_{i=1}^{n} u_q(x_i)^m}. \qquad (16)$$

3. Вычисление новых значений функций принадлежности:

$$u_q(x_i) = \frac{1}{\sum_{t=1}^{c} \left(\frac{d(v_q, x_i)}{d(v_t, x_i)} \right)^{\frac{2}{m-1}}}. \qquad (17)$$

4. Шаги 2 и 3 повторяются до тех пор, пока не будет выполнено заданное число итераций s или не будет достигнута заданная точность $|J(U,V) - J'(U,V)| \le \varepsilon$, где $J(U,V)$, $J'(U,V)$ – значения целевой функции на двух последовательных итерациях.

При применении FCM-алгоритма определяются локально-оптимальное нечеткое разбиение, описываемое совокупностью функций принадлежности, и координаты центров кластеров. Для получения адекватных результатов нечеткой кластеризации необходимо многократное выполнение FCM-алгоритма при заданном числе кластеров для различных исходных нечетких разбиений для принятия окончательного решения об искомой нечеткой кластеризации.

В качестве показателя качества разбиения рекомендуется использовать индекс Се-Бени:

$$XB = \frac{J(U,V)}{n \cdot \min\limits_{k \neq t} d(v_k, v_t)}, \qquad (18)$$

где $J(U,V)$ – целевая функция, вычисляемая по формуле (14).

Индекс Се-Бени учитывает, как нечеткие степени принадлежности объектов центрам кластеров, так и геометрическое расположение центров кластеров и объектов, что в большинстве случаев позволяет получить адекватные результаты кластеризации. В результате работы FCM-алгоритма центры искомых кластеров «стягиваются» к скоплениям объектов, обеспечивая при этом минимальное сходство между признаками объектов, принадлежащих разным кластерам.

Еще один активно используемый показатель качества нечеткой кластеризации – коэффициент *FPC* (fuzzy partition coefficient):

$FPC = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}\left(\mu_j(x_i)\right)^m$, где m – фаззификатор; n – число объектов; c – число кластеров; $\mu_j(x_i)$ – степень принадлежности объекта x_i j-му кластеру.

PAM-алгоритм (partitioning around medoids, k-medoids)

PAM-алгоритм предполагает минимизацию функции:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i, x_j) z_{ij} \to \min,$$

при этом

$$\sum_{i=1}^{n} z_{ij} = 1, \ j = \overline{1,n}.$$

$$z_{ij} \le y_i, \ i = \overline{1,n}; \ j = \overline{1,n}.$$

$$\sum_{i=1}^{n} y_i = k, \text{ где } k - \text{число кластеров.}$$

$$z_{ij}, y_i \in \{0, 1\}, \ i = \overline{1,n}; \ j = \overline{1,n}.$$

$d(x_i, x_j)$ – мера расстояния между объектами *i* и *j*; z_{ij} – переменная, которая гарантирует, что расстояние только между объектами из одного кластера будет вычислено в целевой функции.

Ограничения гарантируют, что:

- каждый объект принадлежит одному и только одному кластеру;
- каждый объект относится к медоиде, представляющей его кластер;
- есть в точности *k* кластеров;
- решающая переменная принимает значения 0 или 1.

PAM-алгоритм имеет две фазы:

Фаза **Build**:

1. Выбрать *k* объектов в качестве медоид.

2. Построить матрицу расстояний, если она не была задана.

3. Отнести каждый объект к ближайшей медоиде.

Фаза **Swap**:

4. Для каждого кластера найти объекты, снижающие среднее расстояние, и если такие объекты есть, выбрать те, которые снижают его сильней всего, в качестве медоид.

5. Если хотя бы одна медоида поменялась, вернуться к шагу 3, иначе завершить алгоритм.

Оценка числа кластеров. Показатели качества кластеризации

В ходе решения задачи кластеризации на каждом из ее этапов возможны ситуации, влияющие на итоговый результат и искажающие его. Эти ситуации могут возникнуть на каждом этапе кластеризации из-за: 1) ошибочно-выделенных характеристик исходных объектов (что часто бывает при задании векторного представления текстов); 2) неверно определенной меры для сравнения объектов; 3) неверного выбора самого алгоритма кластеризации применительно к исходным данным.

Для уверенности в том, что кластерная структура корректна и годится для дальнейшего использования, необходимо, чтобы она отвечала требованиям, вытекающим из постановки задачи кластеризации:

1) *компактность* – объекты одного кластера должны быть как можно ближе друг к другу, что может быть выражено: через расстояния между объектами в кластере, через плотность объектов внутри кластера или же через объем, занимаемый кластером в многомерном пространстве;

2) *отделимость* – расстояние между различными кластерами должно быть как можно больше; что может быть выражено через расстояние между ближайшими объектами разных кластеров или через расстояние между наиболее удаленными друг от друга объектами кластеров или через расстояние между кластерными центрами;

3) *концентрация* (для алгоритмов кластеризации, в которых выделяется центр кластера) – объекты одного кластера должны быть сконцентрированы вокруг центра кластера.

В настоящее время известны различные подходы к оценке качества кластеризации с использованием алгоритмов кластеризации, основанные на применении различных функций, представляющих те или иные показатели качества кластеризации. Более того, для оценки качества четкой кластеризации (когда не допускается пересечение кластеров) и кластеризации в условиях неопределенности (когда допускается пересечение кластеров) используются различные показатели.

Оценка качества четкой кластеризации может быть выполнена с использованием таких показателей качества, как: модифицированная Hubert Γ Statistic или ее нормализованная версия $\hat{\Gamma}$; индекс Calinski – Harabasz; индекс Данна *Dunn*; индекс *DB* (Девида – Болдуина); индексы *CS* или *PS* validity; *SD* (Scatter – Distance), *SDbw* (Scatter – Density), *RMSSTD* (Root Mean Square Standard Deviation), *RS* (R Squared) индексы; индекс оценки си-

луэта (Silhouette Index); индекс MB (Maulik – Bandoypadhyay), индекс плотности $CDbw$ и др.

Считается, что лучшими по точности оценки являются: $CDbw$, индекс оценки силуэта, $Dunn$ и DB. Показано, что не существует универсального решения для оценки качества четкой кластеризации: для любого показателя существует такое множество объектов, на котором его оценка качества неверна; поэтому для повышения эффективности в оценке качества и получения объективного результата лучше использовать не одним показатель качества, а их совокупность.

Для оценки качества нечеткой кластеризации используются такие показатели качества кластеризации, как: показатели разбиения PC и HZ, энтропия разбиения PE, Индекс Fukuyama – Sugeno FS, индекс Се – Бени XB, индекс отделимости-компактности SC, индекс плотности nCS, нечеткий общий гиперобъем FH, средняя плотность нечеткого разбиения $FAPD$, индекс плотности нечеткого разбиения FPD. Одни из них основаны на использовании только степеней принадлежности объектов центрам кластеров, другие учитывают и степени принадлежности объектов центрам кластеров, и геометрические свойства как самих объектов кластеризации, так и центров кластеров.

При решении большинства практических задач с применением алгоритмов нечеткой кластеризации, в случае, когда множество объектов образовано кластерами гиперсферической формы, в качестве показателя качества кластеризации целесообразно использовать индекс Се – Бени и его модификации, обеспечивающие получение адекватных результатов кластеризации и характеризующиеся невысокой вычислительной сложностью. В случае, когда множество объектов образовано кластерами в форме гиперэллипсоидов, в качестве показателей качества кластеризации следует использовать нечеткий общий гиперобъем FH, среднюю плотность разбиения $FAPD$, индекс плотности разбиения FPD и индекс плотности nCS, обеспечивающие получение

адекватных результатов кластеризации, но характеризующиеся более высокой вычислительной сложностью.

Определение числа кластеров можно выполнить, сравнив значения показателя качества кластеризации при различных значениях числа кластеров: искомое число кластеров то, при котором показатель качества достигает своего экстремального значения (минимального или максимального – в зависимости от смысла, заложенного в показатель).

Определение числа кластеров в случае иерархической кластеризации может быть выполнено, например, с использованием метода локтя (elbow) (метода колена (knee)): большее значение величины изменения кластерного расстояния соответствует искомому числу кластеров.

Кроме того, целесообразно использовать индекс кластерного силуэта.

Индекс кластерного силуэта

При вычислении индекса кластерного силуэта определяется значение силуэта для каждого объекта кластеризации, поэтому можно оценить как качество отнесения отдельного объекта кластеру, так и качество кластеризации в целом (рисунок 7).

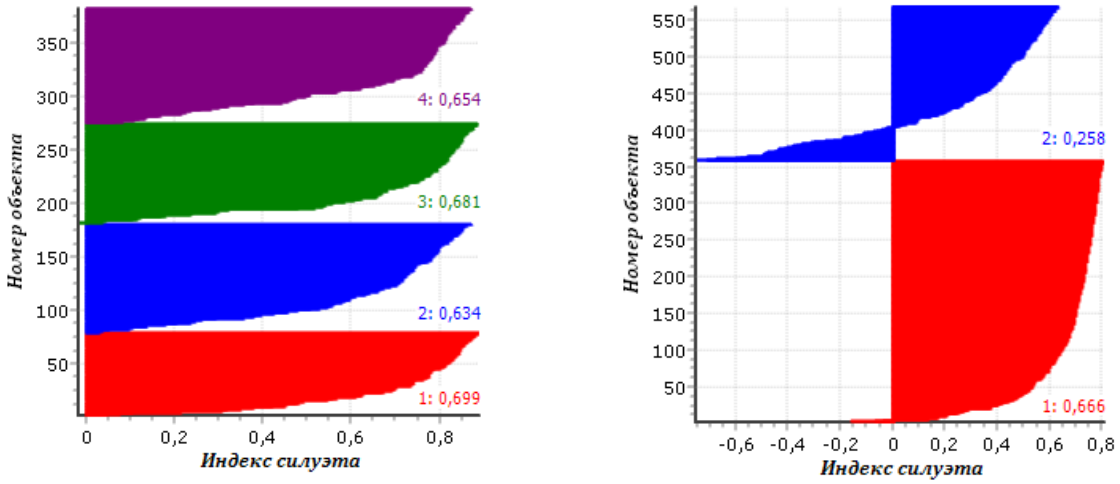


Рисунок 7. График силуэтов: слева – разделение на 4 кластера, справа – разделение на 2 кластера

Пусть в результате кластеризации объект z_i $(i=\overline{1,s})$ отнесен к кластеру c_r. Для определения силуэта для каждого отдельного объекта z_i $(i=\overline{1,s})$ вычисляются следующие значения: $a(z_i)$ – среднее расстояние от объекта z_i до других объектов кластера c_r, которому принадлежит объект z_i; $g_p(z_i)$ – среднее расстояние от объекта z_i до объектов из кластера c_p $(r \neq p=\overline{1,c})$, которому не принадлежит объект z_i; $b(z_i)=\min\limits_{p\neq r} g_p(z_i)$ – показатель различия объекта z_i с объектами ближайшего кластера:

$$a(z_i)=\frac{1}{|c_r|}\sum_{z_j\in c_r} d(z_i,z_j),\ (z_i\in c_r); \tag{19}$$

$$b(z_i)=\min_{p\neq r}\left(g_p(z_i)\right)=\min\left\{\frac{1}{|c_p|}\sum_{z_j\in c_p} d(z_i,z_j), r\neq p=\overline{1,c}\right\}. \tag{20}$$

Тогда индекс оценки силуэта $sil(z_i)$ каждого отдельного объекта z_i $(i=\overline{1,s})$ определяется в соответствии с формулой:

$$sil(z_i)=\frac{a(z_i)-b(z_i)}{\max(a(z_i),\ b(z_i))}. \tag{21}$$

При вычислении расстояния между объектами в (19) и (20) могут использоваться различные меры, но, как правило, самой распространенной является Евклидово расстояние.

Значение индекса оценки силуэта лежит в интервале $[-1;1]$; чем ближе значение индекса оценки силуэта $sil(z_i)$ к единице, тем лучше для объекта z_i был определен кластер принадлежности. Отрицательное значение индекса оценки силуэта $sil(z_i)$ свидетельствует о том, что объект плохо кластеризован (кластер для него определен неверно). Значение индекса оценки силуэта $sil(z_i)$, близкое к нулю, означает, что объект равноправно мог быть отнесен в несколько кластеров. Для кластера, состоящего из одного объекта, индекс оценки силуэта считается равным нулю.

График кластерных силуэтов объединяет графики индекса оценки силуэта $sil(z_i)$, отсортированные по возрастанию значений внутри каждого кластера. На рисунке 5 изображены примеры графиков кластерных силуэтов. По рисунку 5 (слева) можно судить о хорошем качестве разбиения объектов на 4 кластера (лишь у трех объектов небольшие по модулю отрицательные значения индекса оценки силуэта). По рисунку 5 (справа) видно, что первый кластер построен достаточно хорошо, а во втором есть не очень удачно классифицированные объекты.

Оценка силуэта для всей кластерной структуры достигается усреднением индекса оценки силуэта по объектам:

$$SWC = \frac{1}{s}\sum_{i=1}^{s} sil(z_i). \qquad (22)$$

Лучшее разбиение характеризуется максимально близким к единице значением SWC, что достигается, когда расстояние между объектами внутри кластеров мало, а расстояние между объектами соседних кластеров велико.

Трансдуктивное обучение

В отличии от ***индукции***, являющейся рассуждением от ***частного*** (наблюдаемых объектов обучения) к ***общему*** (закономерностям общего характера), **трансдукцией** называют **выводы о частных случаях** (тестовых данных) **на основании частных случаев** (данных обучения).

Различия между этими методами построения выводов особенно интересны, когда прогноз, полученный с помощью трансдуктивной модели, невозможно получить, используя модель индуктивную.

Подобные ситуации возникают, когда в результате трансдутивного вывода на различных тестовых наборах получаются взаимно противоречивые прогнозы.

Понятие трансдукции было ***введено Владимиром Вапником*** в девяностых годах двадцатого века. По мнению Вапника трансдукция может быть отнесена к индукции, поскольку индукция требует решения общей задачи (восстановления функции) перед решением задачи более конкретной (вычисление результатов для новых объектов): «Решая интересующую Вас задачу, не стоит решать более общую задачу на промежуточном шаге. Постарайтесь получить ответ, который Вам действительно нужен, а не более общий.»

Примером обучения, не являющегося индуктивным, может быть случай двоичной (бинарной) классификации, в котором входные данные склонны разделяться на две группы. Большой объём контрольных данных может помочь в поиске кластеров, давая полезную информацию о метках классов. Те же выводы не могут быть достигнуты с помощью модели, восстанавливающей функцию лишь на основании обучающей выборки. Может показаться, что это пример тесно связанного с трансдукцией частичного обучения, но у Вапника была несколько иная мотивация. Примером алгоритма этой категории может послужить трансдуктивная машина опорных векторов (Transductive Support Vector Machine, TSVM).

Третья возможная причина, ведущая к трансдукции, возникает при необходимости в приближении. Если построение точного ответа вычислительно невозможно, то можно по крайней мере попытаться убедиться в том,

что приближения хороши на тестовых данных. В этом случае тестовые данные могут иметь произвольное распределение (необязательно связанное с распределением обучающих данных), что недопустимо в случае частичного обучения. Примером алгоритма, подпадающего под эту категорию, может является Машина Байесовых Комитетов (Bayesian Committee Machine, BCM).

Обучение с частичным привлечением учителя

Обучение с частичным привлечением учителя или полуавтоматическое обучение или частичное обучение (англ. Semi-supervised learning) – способ машинного обучения, разновидность обучения с учителем, которое также использует неразмеченные данные для тренировки – обычно небольшое количество размеченных данных и большое количество неразмеченных данных.

Полуавтоматическое обучение занимает промежуточную позицию между обучением без учителя (без привлечения каких-либо размеченных данных для тренировки) и обучением с учителем (с привлечением лишь размеченных данных).

Многие исследователи машинного обучения обнаружили, что неразмеченные данные, при использовании в сочетании с небольшим количеством размеченных данных, могут значительно улучшить точность обучения.

Задание размеченных данных для задачи обучения часто требует квалифицированного человека (например, для транскрибирования аудио файла) или физического эксперимента (например, для определения 3D структуры белка или выявления наличия нефти в определенном регионе). Поэтому затраты на разметку данных могут сделать процесс обучения с использованием лишь размеченных данных невыполнимым, в то время как процесс задания неразмеченных данных не является очень затратным. В таких ситуациях, полуавтоматическое обучения может иметь большое практическое значение. Такое обучение также представляет интерес в сфере машинного обучения и как модель для человеческого обучения.

Задача обучения

Как и в рамках обучения с учителем, нам дается множество независимых одинаково распределенных примеров с соответствующими метками .

Кроме того, нам дано неразмеченных примеров . Цель полуавтоматической обучения заключается в том, чтобы использовать эту комбинированную информацию для достижения лучших результатов производительности классификации, которую можно получить или путем отбрасывания неразмеченных данных и использование обучения с учителем, или путем отбрасывания меток и использование обучения без учителя.

Полуавтоматическое обучение может принадлежать к трансдуктивному обучению или индуктивному обучению.

Целью трансдуктивного обучения является выведение правильных меток только для неразмеченных данных .

Целью индукции является выведение правильного отображения из в .

Мы можем представлять задачу обучения как экзамен, а размеченные данные – как несколько примеров, которые учитель решил в классе. Учитель также предоставляет набор нерешенных задач.

В постановке трансдуктивного обучения, эти нерешенные задачи является экзаменом, который забирают домой, и вы хотите хорошо его составить в целом.

В постановке индуктивного обучения, эти практические задачи являются подобными тем, с которыми вы столкнетесь на экзамене в классе. Необязательно (и, согласно принципу Вапника, неблагоразумно) проводить трансдуктивное обучение путем логического вывода правила классификации для всех входных данных. Однако, на практике, алгоритмы, формально предназначенные для трансдукции или индукции, часто используются как взаимозаменяемые.

Предположения, которые используются в полуавтоматическом обучении

Для того, чтобы использовать неразмеченные данные, нужно присвоить некоторую структуру для основного распределения данных. Алгоритмы полуавтоматического обучения используют по крайней мере одно из таких предположений.

1. **Предположение плавности**

Точки, которые лежат близко друг от друга, размечены одинаково с большей вероятностью. Такое же предположение в основном используется и в обучении с учителем и дает преимущество в использовании геометрически простых решений. В случае полуавтоматического обучения предположение плавности дополнительно дает преимущество для разграничения в регионах с низкой плотностью, где меньше точек, которые расположены близко друг от друга, но разных классов.

2. **Предположение кластеризованности**

Данные, как правило, образуют дискретные кластеры, и точки из одного кластера размечены одинаково с большей вероятностью (хотя данные, которые используют одинаковые метки, могут быть расположены в нескольких различных кластерах). Это особый случай предположения плавности, который приводит к обучению с использованием алгоритмов кластеризации.

3. **Предположение избыточности данных**

Это предположение применимо, когда измерения данных избыточны, то есть генерируются определенным процессом, имеющим только несколько степеней свободы. В этом случае неразмеченные данные позволяют изучить генерирующий процесс и за счёт этого снизить размерность.

Например, человеческий голос контролируется несколькими голосовыми связками, а изображение различных выражений лица контролируются несколькими мышцами. В этих случаях удобнее использовать генерирующее пространство, чем пространство всех возможных акустических волн или изображений, соответственно.

Предварительное сокращение размерности. Факторный анализ

Факторный анализ – метод, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные

зависят от меньшего количества неизвестных переменных и случайной ошибки.

Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и в то же время компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Две основных задачи факторного анализа:

– определение взаимосвязей между переменными, (классификация переменных);

– сокращение числа переменных необходимых для описания данных.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные. Например, при анализе оценок, полученных по нескольким шкалам, может быть выявлено, что они сходны между собой и имеют высокий коэффициент корреляции. В результате, можно предположить, что существует некоторая латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют фактором. Данный фактор влияет на многочисленные показатели других переменных, что приводит к возможности и необходимости выделить его как наиболее общий, более высокого порядка. Для выявления наиболее значимых факторов и, как следствие, факторной структуры, наиболее оправданно применять метод главных компонент (МГК, PCA, Principal Component Analysis).

Суть данного метода состоит в замене коррелированных компонентов некоррелированными факторами. Другой важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить остальные из анализа, что упрощает интерпретацию результатов. Достоинство МГК также в том, что он – единственный математически обоснованный метод факторного анализа. По утверждению ряда исследователей МГК не является методом факторного анализа, поскольку не расщепляет дисперсию индикаторов на общую и уникальную.

Основной смысл факторного анализа заключается в выделении из всей совокупности переменных только небольшого числа латентных независимых друг от друга группировок, внутри которых переменные связаны сильнее, чем переменные, относящиеся к разным группировкам.

Факторный анализ может быть:

– разведочным, если осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках;

– конфирматорным (подтверждающим), если предназначен для проверки гипотез о числе факторов и их нагрузках.

Условия применения факторного анализа

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

– все признаки должны быть количественными;

– число наблюдений должно быть не менее чем в два раза больше числа переменных;

– выборка должна быть однородна;

– исходные переменные должны быть распределены симметрично;

– факторный анализ осуществляется по коррелирующим переменным.

Два основных понятия факторного анализа: **фактор** – скрытая переменная и **нагрузка** – корреляция между исходной переменной и фактором.

Сущностью факторного анализа является процедура вращения факторов, то есть перераспределения дисперсии по определённому методу. Цель ортогональных вращений – определение простой структуры факторных нагрузок, целью большинства косоугольных вращений является определение простой структуры вторичных факторов, то есть косоугольное вращение следует использовать в частных случаях. Поэтому ортогональное вращение предпочтительнее.

Согласно определению Мюльека простая структура соответствует требованиям:

- в каждой строке матрицы вторичной структуры V должен быть хотя бы один нулевой элемент;
- для каждого столбца k матрицы вторичной структуры V должно существовать подмножество из r линейно-независимых наблюдаемых переменных, корреляции которых с k-м вторичным фактором – нулевые (в итоге каждый столбец матрицы должен содержать не менее r нулей);
- у одного из столбцов каждой пары столбцов матрицы V должно быть несколько нулевых коэффициентов (нагрузок) в тех позициях, где для другого столбца они ненулевые (это требование гарантирует различимость вторичных осей и соответствующих им подпространств размерности $(r – 1)$ в пространстве общих факторов;
- при числе общих факторов больше четырёх в каждой паре столбцов должно быть некоторое число нулевых нагрузок в одних и тех же строках (это требование дает возможность разделить наблюдаемые переменные на отдельные скопления);
- для каждой пары столбцов матрицы V должно быть как можно меньше значительных по величине нагрузок, соответствующих одним и тем же строкам (это требование обеспечивает минимизацию сложности переменных).

В определении Мьюлейка через r обозначено число общих факторов, а V – матрица вторичной структуры, образованная координатами (нагрузками) вторичных факторов, получаемых в результате вращения.

Вращение бывает:

- ортогональным;
- косоугольным.

При первом виде вращения каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, поэтому факторы оказываются независимыми, некоррелированными друг от друга (к этому типу методов факторного анализа относится МГК). Второй вид вращения – это преобразование, при котором факторы коррелируют друг с другом. Преимущество косоугольного вращения состоит в следующем: когда в результате его выполнения получаются ортогональные факторы, можно быть уверенным, что эта ортогональность действительно им свойственна, а не привнесена искусственно.

Наиболее часто использует ортогональный метод вращения «варимакс». Метод «варимакс» максимизирует разброс квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок. В результате простая структура получается для каждого фактора в отдельности.

Главной проблемой факторного анализа является выделение и интерпретация главных факторов. При отборе компонент приходится сталкиваться с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов. Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой.

1. *Критерий Кайзера* или *критерий собственных чисел*. Этот критерий предложен Кайзером, и является наиболее широко используемым. От-

бираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

2. *Критерий каменистой осыпи* или *критерий отсеивания*. Он является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения возможно изобразить в виде простого графика. Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» – «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. Однако этот критерий отличается высокой субъективностью и, в отличие от предыдущего критерия, статистически не обоснован.

Недостатки обоих критериев заключаются в том, что первый иногда сохраняет слишком много факторов, в то время как второй, напротив, может сохранить слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный вопрос: когда полученное решение может быть содержательно интерпретировано. В этой связи предлагается использовать ещё несколько критериев.

3. *Критерий значимости*. Он особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Но критерий непригоден для поиска изменений в модели и реализуем только в факторном анализе по методу наименьших квадратов или максимального правдоподобия.

4. *Критерий доли воспроизводимой дисперсии*. Факторы ранжируются по доле детерминируемой дисперсии, когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса. Недостатки критерия: во-первых, субъективность выделения, во-вторых, специфика данных может

быть такова, что все главные факторы не смогут совокупно объяснить желательного процента разброса. Поэтому главные факторы должны вместе объяснять не меньше 50,1 % дисперсии.

5. *Критерий интерпретируемости и инвариантности*. Данный критерий сочетает статистическую точность с субъективными интересами. Согласно ему, главные факторы можно выделять до тех пор, пока будет возможна их ясная интерпретация. Она, в свою очередь, зависит от величины факторных нагрузок, то есть если в факторе есть хотя бы одна сильная нагрузка, он может быть интерпретирован. Возможен и обратный вариант – если сильные нагрузки имеются, однако интерпретация затруднительна, от этой компоненты предпочтительно отказаться.

Практика показывает, что, если вращение не произвело существенных изменений в структуре факторного пространства, это свидетельствует о его устойчивости и стабильности данных.

Возможны ещё два варианта:

- сильное перераспределение дисперсии – результат выявления латентного фактора;
- очень незначительное изменение (десятые, сотые или тысячные доли нагрузки) или его отсутствие вообще, при этом сильные корреляции может иметь только один фактор, – однофакторное распределение.

Последнее возможно, например, когда на предмет наличия определённого свойства проверяются несколько групп объектов, однако искомое свойство есть только у одной из них.

Факторы имеют две характеристики: объём объясняемой дисперсии и нагрузки. Если рассматривать их с точки зрения геометрической аналогии, то следует отметить, что фактор, лежащий вдоль оси OX, может максимально объяснять 70 % дисперсии (первый главный фактор), фактор, лежащий вдоль оси OУ, способен детерминировать не более 30 % (второй главный фактор). То есть в идеальной ситуации вся дисперсия может быть объяснена двумя главными факторами с указанными долями.

В обычной ситуации может наблюдаться два или более главных факторов. Кроме того, остаётся часть неинтерпретируемой дисперсии (геометрические искажения), исключаемая из анализа по причине незначимости. Нагрузки, опять же с точки зрения геометрии, есть проекции от точек на оси OX и OУ (при трёх- и более факторной структуре также на ось OZ). Проекции – это коэффициенты корреляции, точки – наблюдения, таким образом, факторные нагрузки являются мерами связи.

Так как сильной считается корреляция с коэффициентом Пирсона $R \geq 0{,}7$, то в нагрузках нужно уделять внимание только сильным связям.

Факторные нагрузки могут обладать свойством *биполярности* – наличием положительных и отрицательных показателей в одном факторе. Если биполярность присутствует, то показатели, входящие в состав фактора, дихотомичны и находятся в противоположных координатах.

Метод главных компонент

Метод главных компонент – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации. Предложен Карлом Пирсоном в 1901 году. Применяется во многих областях, в том числе, в эконометрике, биоинформатике, обработке изображений, для сжатия данных и др.

Вычисление главных компонент может быть сведено к вычислению сингулярного разложения матрицы данных или к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных.

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем при-

знаков. Так, например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер-рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показали исследования, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам.

Метод главных компонент является методом линейного снижения размерности данных.

Метод главных компонент считается статистическим методом. Однако есть другой подход, приводящий к методу главных компонент, но не являющийся статистическим. Этот подход связан с получением наилучшей проекции точек наблюдения в пространстве меньшей размерности. Для решения подобной задачи необходимо знать матрицу вторых моментов.

В статистическом подходе задача заключается в выделении линейных комбинаций случайных величин, имеющих максимально возможную дисперсию. Он опирается на ковариационную или корреляционную матрицу этих случайных величин. У этих двух разных подходов есть общий аспект: использование матрицы вторых моментов как исходный для начала анализа.

Математической моделью, на которой основываются методы многомерного статистического анализа (в том числе и методы факторного анализа), является многомерное нормальное распределение.

Из центральной предельной теоремы следует, что предельным распределением одномерных независимых случайных величин является одномерный нормальный закон.

Из обобщённой центральной предельной теоремы получаем, что предельным распределением в случае нескольких измерений является многомерное нормальное распределение.

В настоящее время многомерные методы, основанные на нормальном распределении, нашли широкое распространение при изучении различных процессов в экономике.

Среди математических методов многомерного анализа выделяют следующие.

1. Методы на основе корреляции.

При изучении корреляции рассматриваются различные коэффициенты корреляции.

Выборочные коэффициенты корреляции используются для оценки соответствующих параметров распределения.

Частный коэффициент корреляции измеряет зависимость между случайными величинами, когда действие других коррелированных случайных величин исключено.

При помощи множественного коэффициента корреляции распространяется понятие коэффициента корреляции на измерение зависимости между одной случайной величиной и множеством случайных величин.

2. Аналоги одномерных статистических методов в многомерном анализе.

Многие проблемы, решаемые в многомерном статистическом анализе, когда изучаются многомерные совокупности, имеют свои аналоги при изучении одномерных совокупностей. Для этих проблем выбор системы координат связан с линейным преобразованием переменных.

3. Проблемы системы координат.

В ряде случаев удачный выбор новой системы координат может наиболее экономным способом выявить некоторые важные для исследователя свойства многомерной случайной совокупности.

Примером может служить выявление главных компонент, т.е. отыскание такой нормализованной линейной комбинации случайных величин, чтобы ее дисперсия была максимальной или минимальной. Это равноценно повороту осей, который приводит ковариационную матрицу к диагональной форме. Другой пример – нахождение канонических корреляций. Для решения подобных задач требуется определение характеристических корней различных систем линейных алгебраических уравнений.

4. Проблемы классификации.

Это разбиение множества случайных величин на подмножества. Возникает важный вопрос проверки гипотезы о независимости подмножеств. Факторный анализ, метод главных компонент и кластерный анализ обычно используют в задачах многомерной классификации.

5. Зависимость наблюдений.

Если в экономических исследованиях мы занимаемся анализом временных рядов, то сталкиваемся с наблюдениями над рядами случайных величин, последовательными во времени. Наблюдения в данный момент времени могут зависеть от ранее произведенных наблюдений. Это требует, например, изучения внутрирядной корреляции.

Поскольку в качестве основной статистической модели выступает многомерное нормальное распределение, стоит остановится более подробно на этом распределении, которое полностью распределяется своей квадратичной формой, а последняя зависит от вектора математических ожиданий и ковариационной матрицы. Эта зависимость четко определяется следующей теоремой.

Теорема 1. Если даны вектор μ и положительно определенная матрица Σ, то существует такая многомерная нормальная плотность распределения вероятностей:

$$f(x) = 1/((2\pi)^{\frac{n}{2}} \left|\Sigma\right|^{\frac{1}{2}})\ e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \qquad (23)$$

что математическое ожидание случайного вектора x с этой плотностью распределения есть μ и ковариационная матрица есть Σ.

В данном распределении интересует нас структура ковариационной матрицы и ее связь с корреляционной матрицей. Это можно сделать в общем виде для случайного вектора n-го порядка. Однако удобней обратиться к простейшему многомерному распределению – двумерному.

При рассмотрении двумерного нормального распределения можно легко убедиться в том, что коэффициенты корреляции и дисперсии случайных величин являются основными числовыми характеристиками наряду с математическими ожиданиями. Если конечное число случайных величин n=2, то роль дисперсий выполняет ковариационная (корреляционная) матрица. Элементы этой матрицы получаются из экспериментальных или статистических данных и являются статистическими величинами, требующими своей оценки.

Главные компоненты являются характеристическими векторами ковариационной матрицы.

Множество главных компонент представляет собой удобную систему координат, а соответствующие дисперсии главных компонент характеризуют их статистические свойства. Из общего числа главных компонент для исследования, как правило, оставляют m ($m<n$) наиболее весомых, т.е. вносящих максимальный вклад в объясняемую часть общей дисперсии. Эксперименты показывают, что $m \approx (0{,}1+0{,}25)n$.

Метод главных компонент одинаково хорошо приближает ковариации и дисперсии. Следует отметить еще одно существенное свойство метода – это его линейность и аддитивность.

Векторы главных компонент могут быть найдены как решения однотипных задач оптимизации.

1. Централизуются данные (вычитанием среднего): $x_i = x_i - \bar{x}$. В итоге: $\sum_{i=1}^{m} x_i = 0$
2. Отыскивается первая главная компонента как решение задачи:

   $a_1 = \arg\min_{\|a_1\|=1}(\sum_{i=1}^{m}\|x_i - a_1(a_1, x_i)\|^2)$. Если решение не единственное, выбирается любое.

3. Из данных вычитается проекция на первую главную компоненту:

   $x_i = x_i - a_1(a_1, x_i)$.

4. Отыскивается вторая главная компонента как решение задачи:

   $a_2 = \arg\min_{\|a_2\|=1}(\sum_{i=1}^{m}\|x_i - a_2(a_2, x_i)\|^2)$. Если решение не единственное, выбирается любое.

Процесс продолжается: на шаге $(2k - 1)$ вычитается проекция на $(k - 1)$-ю главную компоненту (к этому моменту проекции на предшествующие $(k - 2)$ главные компоненты уже вычтены):

$x_i = x_i - a_{k-1}(a_{k-1}, x_i)$

и на шаге $2k$ определяется k-я главная компонента как решение задачи:

$a_k = \arg\min_{\|a_k\|=1}(\sum_{i=1}^{m}\|x_i - a_k(a_k, x_i)\|^2)$. Если решение не единственное, выбирается любое.

На каждом подготовительном шаге $(2k - 1)$ вычитается проекция на предшествующую главную компоненту.

Найденные векторы $\{a_1, \ldots, a_{n-1}\}$ ортонормированы просто в результате решения описанной задачи оптимизации, однако, чтобы не дать ошибкам вычисления нарушить взаимную ортогональность векторов главных компонент, можно включать требование $a_k \perp \{a_1, \ldots, a_{k-1}\}$ в условия задачи оптимизации.

Неединственность в определении a_k помимо тривиального произвола в выборе знака может быть более существенной и происходить, например, из

условий симметрии данных. Последняя главная компонента a_n – единичный вектор, ортогональный всем предыдущим a_k.

Сложности и проблемы, которые могут возникнуть при применении кластерного анализа

Постановка задачи кластеризации сложна и неоднозначна, так как:

– число кластеров в общем случае неизвестно;

– выбор меры «похожести» или близости свойств объектов между собой, как и критерия качества кластеризации, всегда носит субъективный характер.

Разные алгоритмы кластеризации могут давать различные результаты кластеризации (относительно некоторой части данных).

Сравнительный анализ иерархических и итеративных алгоритмов кластеризации

Обычно алгоритмы иерархической кластеризации применяются при небольшом числе объектов. Такие алгоритмы удобно использовать для визуализации результатов кластеризации в виде дендрограмм.

Итеративные алгоритмы кластеризации обычно используют при большом числе объектов.