Глава 19

# КЛАССИФИКАЦИЯ

Навостри ум свой математикой, если не найдешь для этого никакого иного средства, остерегайся только классификации букашек, поверхностное знание которой совершенно бесполезно, а точное уводит в бесконечность. И не забывай, что число фибр твоего мозга, их складок и извилин конечно. Там, где сидит какая-нибудь история бабочки, нашлось бы, может быть, место для биографий Плутарха, которые могли бы вдохновить тебя.

*Г. К. Лихтенберг*

При статистическом анализе таблицы данных, состоящей из нескольких столбцов (признаков), необходимо иметь в виду *эффект существенной многомерности*, из-за которого к верным выводам можно прийти лишь при одновременном учете всей совокупности взаимосвязанных признаков. Так, точки и кружки на рис. 1 почти не отличаются друг от друга по каждой из координат в отдельности, но очевидным образом разделяются по новому признаку — сумме координат.

Похожий случай приводится в [1, с. 15]: попытка различить два типа потребительского поведения семей сначала по одному признаку (расходы на питание), потом по другому (расходы на промышленные товары и услуги) не дала результата, в то время как одновременный учет обоих признаков позволил обнаружить значимое различие между анализируемыми совокупностями семей. (См. также пример 1 в гл. 23.)

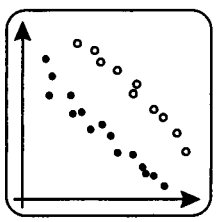


**Рис. 1**

Рассмотрим еще один пример, показывающий, что удачная классификация может даже привести к появлению нового направления исследований (см. [4, с. 35]).

«С давних пор астрономы знали о различной светимости звезд, т. е. о различной их «истинной яркости». В конце XIX в. были открыты также различные спектральные классы звезд, попросту говоря — различный цвет их излучения (от красного до голубого). До 1913 г. эти характеристики существовали в представлении ученых раздельно, но вот (независимо друг от друга) датский астроном Герцшпрунг и американец Ресселл сопоставили их между собой и построили двумерную проекцию объектов-звезд на плоскость признаков спектр — светимость. Результаты оказались неожиданными (рис. 2).

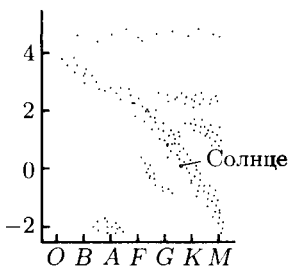



**Рис. 2**

Астрономы увидели, что звезды не распределены в пространстве этих признаков равномерно, а образуют несколько ярко выраженных кластеров, причем стало возможным предсказать эволюцию звезд по значениям их основных характеристик. С тех пор диаграмма Герцшпрунга—Ресселла стала одним из важных инструментов в работе современных астрономов.»

Если число признаков $m > 3$, то разбиение множества объектов на компактные группы (так называемые *кластеры*)*) может оказаться непростой задачей. Данная глава посвящена знакомству с некоторыми подходами к ее решению.

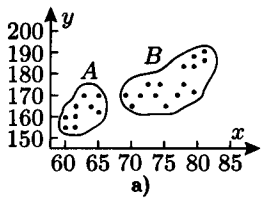
## § 1. НОРМИРОВКА, РАССТОЯНИЯ И КЛАССЫ

Статистическая однородность — понятие, базисное для статистики; общепринято, что какую-либо обработку статистических данных (усреднение, установление связей и т. д.) надо производить только в однородных группах наблюдений.

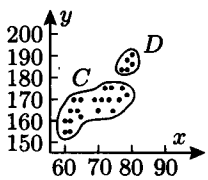
*И. Д. Мандель, «Кластерный анализ»*

Разбиение объектов на классы может в значительной степени *зависеть от выбора единиц измерения* (масштабов шкал) признаков: килограммы или фунты, сантиметры или дюймы.

**Пример 1 ([52, с. 26]).** Студенты группы записывают свои вес ($x$) и рост ($y$). По этим данным на плоскости строится *диаграмма рассеяния* («облако» точек с координатами $(x_i, y_i)$, $i = 1, \ldots, n$, где $n$ — число студентов в группе). Масштабы по осям задаются произвольно. На рис. 3, *а* девушки ($A$) довольно четко отделяются от юношей ($B$). На рис. 3, *б* шкала на оси веса сжата вдвое. При этом более естественным представляется уже деление на высоких юношей ($D$) и всех остальных студентов ($C$).



а)



б)

Рис. 3

В приведенном примере при классификации не следует считать расстоянием между объектами евклидово расстояние между соответствующими точками $(x_i, y_i)$ на плоскости, так как признаки имеют разные единицы измерения. Требуется предварительная нормировка показателей, переводящая их в безразмерные величины. Перечислим наиболее распространенные типы нормировки одномерных наблюдений $Z_1, \ldots, Z_n$.

### Типы нормировки

**N1)** $Z_i' = (Z_i - Z_{min})/(Z_{max} - Z_{min})$.

**N2)** $Z_i' = (Z_i - \overline{Z})/S$, где $\overline{Z} = \frac{1}{n}\sum Z_i$ — среднее арифметическое,

$S^2 = \frac{1}{n}\sum (Z_i - \overline{Z})^2$ — выборочная дисперсия.

**N3)** $Z_i' = (Z_i - MED)/MAD$, где $MED$ — выборочная медиана (см. § 2 гл. 7), $MAD$**) — *(нормированная) медиана абсолютных отклонений* от $MED$:

$$MAD = \frac{1}{\Phi^{-1}(3/4)}\, MED\{|Z_i - MED|,\ i = 1, \ldots, n\},$$

где $\Phi^{-1}(x)$ — функция, обратная к функции распределения закона $\mathcal{N}(0,1)$.***) Такое преобразование менее подвержено влиянию выделяющихся значений $Z_i$.

*) Разные варианты уточнения этого понятия приведены ниже.

**) Сокращение $MAD$ происходит от английского наименования *Median of Absolute Deviations*.

***) Множитель $1/\Phi^{-1}(3/4) \approx 1{,}483$ обеспечивает для выборки из закона $\mathcal{N}(\mu, \sigma^2)$ сходимость $MAD \xrightarrow{d} \sigma$ при $n \to \infty$ (см. П5).

Помимо типа нормировки решающее влияние на результат классификации оказывает *выбор меры близости* между $m$-мерными точками. Приведем основные способы задания расстояния $d_{ij}$ от точки $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ до точки $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$.

### Расстояния между объектами

**D1)** *Метрика города* (рис. 4)*): $d_{ij} = \sum\limits_{l=1}^{m} |x_{il} - x_{jl}|$. При использовании метрики города хорошо выделяются классы, имеющие вид «облака», вытянутого вдоль оси некоторого признака.

В случае, когда координаты объектов принимают только значения 0 и 1, это расстояние равно количеству несовпадающих координат, т. е. длине пути по ребрам единичного $m$-мерного куба из одной вершины в другую (*метрика Хемминга*).

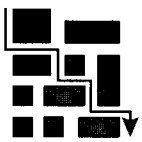


Рис. 4

**D2)** *Евклидова метрика*: $d_{ij} = \left( \sum\limits_{l=1}^{m} (x_{il} - x_{jl})^2 \right)^{1/2}$.

**D3)** *Метрика Чебышёва*: $d_{ij} = \max\limits_{1 \leqslant l \leqslant m} |x_{il} - x_{jl}|$.

Все три расстояния являются частными случаями (соответственно при $p = 1$, 2 и $\infty$) так называемого *расстояния Минковского*

$$d_{ij} = \left( \sum_{l=1}^{m} |x_{il} - x_{jl}|^p \right)^{1/p}.$$

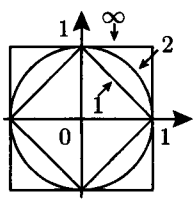



Рис. 5

Известно (см. [90, с. 208]), что при любом $p \geqslant 1$ для расстояния Минковского выполняется *неравенство треугольника*: $d_{ij} \leqslant d_{ik} + d_{kj}$. На рис. 5 изображены единичные «шары» $B_p^m$ для $p = 1$, 2, $\infty$ при $m = 2$. Отношение объемов $\rho_m = V(B_1^m)/V(B_\infty^m) = 1/m!$ при увеличении размерности быстро уменьшается: $\rho_2 = 1/2$, $\rho_5 = 1/120$, $\rho_{10} = 1/3628800 \approx 3 \cdot 10^{-7}$.

**Вопрос 1.** Почему метрика Чебышёва отвечает значению $p = \infty$?

Иногда матрица расстояний (мер близости) $d_{ij}$ между объектами задается непосредственно: например, как таблица экспертных оценок сходства объектов или как матрица прямых измерений близости (скажем, размеров межотраслевых поставок). В этом случае снимается проблема выбора типа нормировки и расстояния. (Однако, заметим, что для некоторых из рассматриваемых ниже методов классификации требуются сами координаты объектов, а не только расстояния $d_{ij}$ между ними.)

На основе заданного расстояния между объектами можно уточнить, какие множества называются *группами однородных объектов* или *классами*. Выделим некоторые

### Типы классов

**C1)** Класс типа ядра [60] (в [56, с. 235] такой класс называется *сгущением*). Все расстояния между объектами внутри

*) Ее также называют *метрикой city-block* или *манхеттенской*.

класса меньше любого из расстояний между объектами класса и остальной частью множества объектов. На рис. 6 сгущениями являются $A$ и $B$. Остальные пары множеств не разделяются с помощью этого определения.

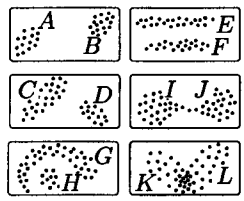


**Рис. 6**

**C2)** Кластер (*сгущение в среднем* [56]). Среднее расстояние внутри класса меньше среднего расстояния объектов класса до всех остальных. Множества $C$ и $D$ теперь разделяются, но у $E$ ($G$) среднее внутреннее расстояние больше, чем среднее расстояние между $E$ и $F$ ($G$ и $H$).

**C3)** Класс типа ленты [60] (*слабое сгущение* [56]). Существует $\tau > 0$ такое, что для любого $\boldsymbol{x}_i$ из класса $S$ найдется такой объект $\boldsymbol{x}_j \in S$, что $d_{ij} \leqslant \tau$, а для всех $\boldsymbol{x}_k \notin S$ справедливо неравенство $d_{ik} > \tau$. В смысле этого определения на рис. 6 разделяются все пары множеств кроме $I$ и $J$, $K$ и $L$.

**C4)** Класс с центром. Существует порог $R > 0$ и некоторая точка $\boldsymbol{x}^*$ в пространстве, занимаемом объектами класса $S$ (в частности, элемент этого множества) такие, что все объекты из $S$ и только они содержатся в шаре радиуса $R$ с центром в $\boldsymbol{x}^*$. Часто в качестве $\boldsymbol{x}^*$ выступает центр масс класса $S$, т. е. координаты центра определяются как средние значения признаков у объектов класса. Множества $I$ и $J$ являются классами с центром, а $E$, $F$ и $G$ — нет.

Обратим внимание на то, что накладывающиеся множества $K$ и $L$ не разделяются при помощи перечисленных определений классов. Тем не менее, в примере 4 из § 5 предлагается способ проведения разделяющей границы между подобными множествами на основе статистической модели случайного выбора из одной из $k$ многомерных нормальных совокупностей.

Вообще, классификация (кластер-анализ) отличается от других разделов статистики большой *зависимостью* результатов расчетов от содержательных установок исследователя.

Научное исследование — это искусство, а правила в искусстве, если они слишком жестки, приносят больше вреда, чем пользы.
*Дж. Томсон, «Дух науки»*

## § 2. ЭВРИСТИЧЕСКИЕ МЕТОДЫ

Подавляющая часть классификаций на практике проводится именно эвристическими методами. Это объясняется относительной простотой и содержательной ясностью таких алгоритмов, возможностью вмешательства в их работу путем изменения одного или нескольких параметров, смысл которых обычно понятен, и невысокой трудоемкостью алгоритмов.

**A1)** Связные компоненты. Все объекты разбиваются на классы *типа ленты, или слабого сгущения* (тип C3 в § 1), где задаваемый параметр $\tau \in (\min d_{ij}, \max d_{ij})$. В этой постановке задача классификации эквивалентна нахождению связных компонент

графа (вершины графа $i$ и $j$ соединены ребром, если $d_{ij} \leqslant \tau$). (Алгоритм выделения связных компонент методом *поиска в глубину* излагается в § 9.) Для выбора величины $\tau$ полезно построить *гистограмму межобъектных расстояний* (высота прямоугольника над промежутком $\Delta_l$ на рис. 7 пропорциональна количеству $d_{ij}$ в $\Delta_l$). При хорошей структурированности данных гистограмма, как правило, имеет два выделяющихся максимума: при $d_{ij} \approx d_{int}$ (*типичное внутриклассовое расстояние*) и при $d_{ij} \approx d_{out}$ (*типичное межклассовое расстояние*). Часто удачным выбором $\tau$ оказывается значение $(d_{int} + d_{out})/2$.

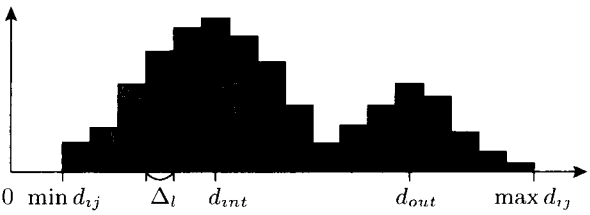



**Рис. 7**

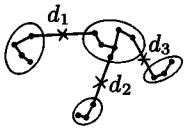



**Рис. 8**

**А2)** Кратчайший незамкнутый путь (КНП). Его также называют *минимальным покрывающим деревом* или *каркасом*. Соединяются ребром две ближайшие точки, затем среди оставшихся отыскивается точка, ближайшая к любой из уже соединенных точек, и присоединяется к ним и т. д. до исчерпания всех точек. Р. Прим в 1957 г. доказал, что построенный таким способом граф имеет минимальную общую длину ребер среди всевозможных соединений, связывающих все вершины (см. [28, с. 60]).

В найденном КНП затем отбрасывают $k-1$ самых длинных дуг и получают $k$ классов (рис. 8).*) Метод позволяет выделять классы произвольной формы.

Заметим, что если разрешается добавлять в граф новые вершины (точки Штейнера), то можно добиться меньшей, чем у КНП, длины пути. Уже для трех точек $A$, $B$ и $C$ на плоскости таких, что наибольший из углов $\triangle ABC$ меньше 120°, существует точка $O$ внутри треугольника, для которой минимальна сумма $|OA| + |OB| + |OC|$.

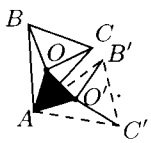



**Рис. 9**

Доказательство [64, с. 77]. Пусть $O$ — некоторая точка внутри $\triangle ABC$. При повороте на 60° вокруг вершины $A$ точки $O$, $B$ и $C$ перейдут в точки $O'$, $B'$ и $C'$, соответственно (рис. 9). Так как $\triangle AOO'$ равносторонний, то $|AO| = |OO'|$. Отрезок $OC$ переходит в отрезок $O'C'$. Поэтому $|OC| = |O'C'|$. Таким образом, сумма $|OA| + |OB| + |OC| = |OO'| + |OB| + |O'C'|$, т. е. равна длине ломаной $BOO'C'$. Она минимальна, когда ломаная является отрезком. Поскольку $\angle AOO' = 60°$, для этого необходимо, чтобы $\angle BOA = 120°$. Другими словами, сторона $AB$ должна быть видна из оптимальной

*) Число $k$ задает исследователь. На практике обычно $2 \leqslant k \leqslant 5$.

точки под углом 120°. Из симметрии то же самое должно быть верно и для двух других сторон $\triangle ABC$. ■

**Вопрос 2.** Как построить оптимальную точку $O$ при помощи циркуля и линейки?

(Постройте на стороне $BC$, как на основании, равносторонний треугольник и опишите вокруг него окружность.)

**Вопрос 3.** Где на плоскости находится точка, сумма расстояний от которой до вершин выпуклого четырехугольника минимальна?

Для произвольного расположения вершин графа на плоскости и любого числа точек Штейнера эффективные алгоритмы построения кратчайшего соединения неизвестны (согласно [28, с. 63]).

**А3)** Метод $k$-средних*) предназначен для выделения классов типа С4 («класс с центром»). Приведем **два варианта**:

**(a)** *Алгоритм Г. Болла и Д. Холла* (1965 г., см. [52, с. 110]). Случайно выбираются $k$ объектов (*эталонов*); каждый объект присоединяется к ближайшему эталону (тем самым образуются $k$ классов); в качестве новых эталонов принимаются центры масс классов.**) После пересчета объекты снова распределяются по ближайшим эталонам и т. д. Критерием окончания алгоритма служит стабилизация центров масс всех классов.

Вместо случайно выбираемых эталонов лучше использовать $k$ наиболее удаленных объектов: сначала отыскиваются два самых удаленных друг от друга объекта, затем $l$-й эталон ($l = 3, \ldots, k$) определяется как наиболее удаленный в среднем от уже имеющихся.

**(b)** *Алгоритм Дж. Мак-Кина* (1967 г., см. [52, с. 98]). Он отличается от метода Болла и Холла тем, что при просмотре списка объектов пересчет центра масс класса происходит после присоединения к нему каждого очередного объекта.

Отметим, что алгоритм Мак-Кина связан с функционалом качества разбиения $F2$ из § 5.

**А4)** Алгоритм «Форель». Случайный объект объявляется центром класса; все объекты, находящиеся от него на расстоянии не большем $R$, входят в первый класс. В нем определяется центр масс, который объявляется новым центром класса и т. д. до стабилизации центра. Затем все объекты, попавшие в первый класс изымаются, и процедура повторяется с новым случайным центром.

«*Форель*»: Первоначальное название ФОРЭЛ — ФОРмальный ЭЛемент. Предложен В. Н. Елкиной и Н. Г. Загоруйко в 1966 г., см. [1, с. 222].

Можно скомбинировать алгоритмы А4 и А2 ([52, с. 67)]: при небольшом $R$ по алгоритму А4 находят $k' > k$ классов; их центры соединяют КНП, из которого удаляют $k - 1$ самых длинных ребер и получают $k$ классов. При этом образуются классы более сложной формы, чем $m$-мерные шары (рис. 10). Здесь важна идея двухэтапности классификации: сначала выделить заведомо компактные маленькие группы, затем произвести их объединение. Так можно успешно классифицировать довольно большие массивы информации (сотни объектов).

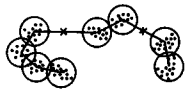


Рис. 10

**А5)** Метод потенциальных ям. Предположим, что каждый объект $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ создает вокруг себя поле притяжения

*) Это название, ставшее популярным, введено Дж. Мак-Кином.

**) Считается, что каждому объекту приписана масса 1.

с некоторой весовой функцией, например, гладким *квартическим ядром*

$$W_i(\boldsymbol{x}) = \bigl[1 - (r_i/R)^2\bigr]^2 I_{\{r_i \leqslant R\}},$$

где $r_i = |\boldsymbol{x} - \boldsymbol{x}_i|$, а параметр $R > 0$ задает эффективный размер области притяжения. Все вместе объекты создают потенциальное поле $U(\boldsymbol{x}) = -\sum W_i(\boldsymbol{x})$. Классам соответствуют потенциальные ямы: объект $\boldsymbol{x}_i$ относится к яме, в которую он «скатывается» при свободном движении. Практически приходится, стартовав с $\boldsymbol{x}_i$, запускать некоторый алгоритм (локальной) минимизации.

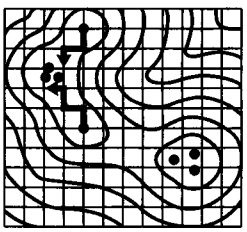


Рис. 11

Одним из простейших методов минимизации функции $m$ переменных $U(\boldsymbol{x})$ является циклический перебор $2m$ точек, соседних с текущей по осям с шагом $\pm h$ ($h$ — точность поиска точки минимума). Если значение $U(\boldsymbol{x})$ в какой-то из них меньше, чем в текущей, происходит перемещение в нее (противоположную по этой оси точку можно не проверять) (рис. 11). Для ускорения полезно запоминать для каждой оси знак успешного перемещения по ней во время предыдущего цикла и сначала пробовать сдвигаться в том же направлении. Критерием окончания служит отсутствие перемещений за время цикла.

Обозначим через $\boldsymbol{y}_i$ конечную точку пути из $\boldsymbol{x}_i$. Объекты $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$ отнесем к одному классу, если $\sum\limits_{l=1}^{m} |y_{il} - y_{jl}| \leqslant 2kh$ (т. е. $\boldsymbol{y}_i$ и $\boldsymbol{y}_j$ близки в *метрике города* D1 из § 1). При этом число перемещений при поиске локального минимума (длина пути) характеризует удаленность $\boldsymbol{x}_i$ от «центра» класса (дна ямы).

Из-за необходимости проведения численной минимизации для каждого $\boldsymbol{x}_i$ метод рекомендуется применять для классификации небольшого числа (нескольких десятков) объектов.

## § 3. ИЕРАРХИЧЕСКИЕ ПРОЦЕДУРЫ

Мне завещал отец:
Во-первых, угождать всем людям без изъятья;
Хозяину, где доведется жить,
Начальнику, с кем буду я служить,
Слуге его, который чистит платья,
Швейцару, дворнику, для избежанья зла,
Собаке дворника, чтоб ласкова была.
*Молчалин в «Горе от ума» А. С. Грибоедова*

Общая схема этих процедур такова: сначала каждый объект считается отдельным классом; на первом шаге объединяются два ближайших объекта, которые образуют новый класс (если сразу несколько объектов (классов) одинаково близки, то выбирается одна случайная пара); вычисляются *меры отдаленности* $\rho$ (см. ниже)*) от этого класса до всех остальных классов, и размерность матрицы межклассовых мер отдаленности сокращается на единицу; шаги процедуры повторяются до тех пор, пока все объекты не объединятся в один класс.

*) Мы не называем их расстояниями из-за того, что не для всех мер отдаленности выполняется неравенство треугольника.

Наиболее известны следующие две процедуры (рис. 12):

**P1)** Метод «ближнего соседа»: $\rho_{min} = \min\limits_{x_i \in S_k,\, x_j \in S_l} d_{ij}$,

**P2)** Метод «дальнего соседа»: $\rho_{max} = \max\limits_{x_i \in S_k,\, x_j \in S_l} d_{ij}$.

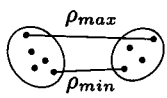
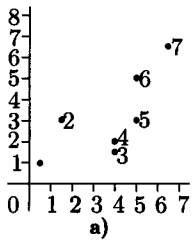



**Рис. 12**

Решение о том, какое из разбиений на классы, получаемых при проведении иерархической процедуры, наиболее содержательно, принимается на основе анализа так называемой *дендрограммы* : по горизонтали откладываются номера объектов, а по вертикали — значения мер отдаленности $\rho(S_k,S_l)$, при которых происходили объединения классов $S_k$ и $S_l$.

Dendron (*греч.*) — дерево.

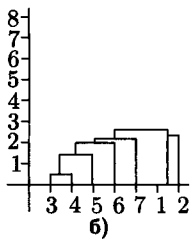
На основе данных, представленных на рис. 13, *а*, построены дендрограммы для процедур P1 (рис. 13, *б*) и P2 (рис. 13, *в*). Хорошо видна следующая особенность метода «ближнего соседа» — *цепочечный эффект*: независимо от формы кластера к нему присоединяются ближайшие к границе объекты. Метод «дальнего соседа» не приводит к подобному эффекту. Подробнее о сравнении разных методов классификации см. в § 7.




а)

Рассмотрим некоторые другие иерархической процедуры.

**P3)** Метод средней связи: $\rho_{ave} = \dfrac{1}{n_k n_l} \sum\limits_{x_i \in S_k} \sum\limits_{x_j \in S_l} d_{ij}$ (здесь $n_k$ и $n_l$ — количества объектов в классах $S_k$ и $S_l$).




б)

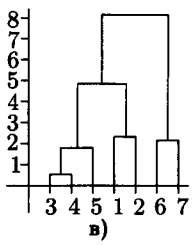
А. Н. Колмогоровым было предложено изящное обобщение метода P3, основанное на понятии *степенного среднего* чисел $c_1 > 0, \ldots, c_n > 0$

$$\overline{c}_\tau = \left( \frac{1}{n} \sum_{i=1}^{n} c_i^\tau \right)^{1/\tau}. \qquad (1)$$

Очевидно, $\overline{c}_1 = \dfrac{1}{n} \sum c_i$ — среднее арифметическое. Устремляя $\tau$ к $+\infty$, $-\infty$ и 0, получаем, соответственно, $\overline{c}_{+\infty} = \max c_i$, $\overline{c}_{-\infty} = \min c_i$, $\overline{c}_0 = (\prod c_i)^{1/n}$ — среднее геометрическое (проверьте!). Таким образом, *мера отдаленности Колмогорова*

$$\rho_K(\tau) = \left[ \frac{1}{n_k n_l} \sum_{x_i \in S_k} \sum_{x_j \in S_l} d_{ij}^\tau \right]^{1/\tau} \qquad (2)$$

включает в себя в качестве частных случаев $\rho_{min}$, $\rho_{max}$ и $\rho_{ave}$.




в)

**Рис. 13**

Average (*англ.*) — средний.

**P4)** Метод центров масс: $\rho_{center} = |\overline{x}_k - \overline{x}_l|^2$, где $\overline{x}_k$ и $\overline{x}_l$ обозначают центры масс $k$-го и $l$-го классов.

Недостатком этого метода является возможность появления *инверсий* — нарушений монотонности увеличения уровня при построении дендрограммы (т. е. объединение классов на некотором шаге процедуры осуществляется при более низком значении меры отдаленности, чем на более раннем шаге).

**Вопрос 4.**
Может ли возникнуть инверсия при классификации методом P4 трех объектов?

**P5)** Метод Уорда*): $\rho_W = \dfrac{n_k n_l}{n_k + n_l} |\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_l|^2$. Как вытекает из приведенной ниже теоремы 1, множитель $\dfrac{n_k n_l}{n_k + n_l}$ позволяет предотвратить появление инверсий.

**Замечание 1.** Покажем, что верно равенство

$$\rho_W = \sum_{\boldsymbol{x}_i \in S_k \cup S_l} |\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{k,l}|^2 - \sum_{\boldsymbol{x}_i \in S_k} |\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k|^2 - \sum_{\boldsymbol{x}_i \in S_l} |\boldsymbol{x}_i - \overline{\boldsymbol{x}}_l|^2, \qquad (3)$$

где $\overline{\boldsymbol{x}}_{k,l}$ — центр масс $S_k \cup S_l$. Таким образом, $\rho_W$ представляет собой *прирост общей внутриклассовой инерции* $V_{int}$ (см. формулу (6) в § 5) *при замене классов* $S_k$ *и* $S_l$ *на их объединение.*

Доказательство. В силу теоремы Гюйгенса (см. решение задачи 3 гл. 16) правая часть (3) равна $n_k|\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_{k,l}|^2 + n_l|\overline{\boldsymbol{x}}_l - \overline{\boldsymbol{x}}_{k,l}|^2$. Сгруппируем массы $S_k$ и $S_l$, поместив их в точки $\overline{\boldsymbol{x}}_k$ и $\overline{\boldsymbol{x}}_l$ соответственно. Общий центр масс $\overline{\boldsymbol{x}}_{k,l}$ при группировке масс не меняется (убедитесь!). Для завершения доказательства остается воспользоваться теоремой о межточечных расстояниях из решения задачи 5 гл. 16. ■

Для вычисления (на очередном шаге иерархической процедуры) мер отдаленности между вновь образованным классом и всеми другими оставшимися классами удобно воспользоваться общей для методов P1—P5 **формулой Г. Ланса и У. Уильямса**:

$$\rho(S_0, S_1 \cup S_2) = C_1\rho_{01} + C_2\rho_{02} + C_3\rho_{12} + C_4|\rho_{01} - \rho_{02}|, \qquad (4)$$

где $\rho_{01} = \rho(S_0, S_1)$ и т. п. Коэффициенты $C_1$—$C_4$ для процедур P1—P5 приведены в следующей таблице:

| Номер | Название | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|---|
| P1 | Ближний сосед | $1/2$ | $1/2$ | $0$ | $-1/2$ |
| P2 | Дальний сосед | $1/2$ | $1/2$ | $0$ | $1/2$ |
| P3 | Средняя связь | $\dfrac{n_1}{n_1+n_2}$ | $\dfrac{n_2}{n_1+n_2}$ | $0$ | $0$ |
| P4 | Центры масс | $\dfrac{n_1}{n_1+n_2}$ | $\dfrac{n_2}{n_1+n_2}$ | $-\dfrac{n_1 n_2}{(n_1+n_2)^2}$ | $0$ |
| P5 | Метод Уорда | $\dfrac{n_0+n_1}{n_0+n_1+n_2}$ | $\dfrac{n_0+n_2}{n_0+n_1+n_2}$ | $-\dfrac{n_0}{n_0+n_1+n_2}$ | $0$ |

Для методов P1 и P2 формула (4) вытекает из тождеств

$$\min\{\alpha,\beta\} = \frac{1}{2}(\alpha+\beta) - \frac{1}{2}|\alpha-\beta|, \quad \max\{\alpha,\beta\} = \frac{1}{2}(\alpha+\beta) + \frac{1}{2}|\alpha-\beta|.$$

Для процедур P3—P5 формула (4) выводится в задаче 1.

*) Предложен Дж. Уордом (J. Ward) в 1963 г.

внутриклассовых связей и (взвешенной) мерой концентрации $\bar{z}_1$ (см. формулу (21)).

Разбиение, минимизирующее $H(S)$, обладает тем свойством, что получаемые классы $S_l$ оказываются *кластерами* (сгущениями в среднем) в смысле определения С2 из § 1: средняя связь $a(S_l) = \sum_{x_i, x_j \in S_l} a_{ij}/n_l^2$ не меньше, чем средняя связь вовне $a(S_l, \overline{S}_l)$ и средняя связь между любыми классами $a(S_l, S_m)$. Точнее, для любых $l \neq m$ имеем $a(S_l, \overline{S}_l)$, $a(S_l, S_m) \leqslant \rho \leqslant a(S_l)$ (см. [1, с. 174]).

Это оптимальное разбиение может быть найдено с помощью иерархической процедуры, на очередном шаге которой объединяются те классы $S_l$ и $S_m$, для которых сумма связей $\sum_{x_i \in S_l, x_j \in S_m} (a_{ij} - \rho)$ максимальна и положительна. Процесс объединения заканчивается, когда такие суммы для всех $l \neq m$ становятся неположительными.

Смысл параметров $\mu$ и $\sigma$ проясняет минимизация по ним $\Delta(S, \mu, \sigma)$ при фиксированном разбиении $S$. Вычислив частные производные по $\mu$ и $\sigma$ функции

$$f(\mu, \sigma) = \sum_{(i,j):\, b_{ij}=1} (a_{ij} - \mu - \sigma)^2 + \sum_{(i,j):\, b_{ij}=0} (a_{ij} - \mu)^2,$$

получим систему уравнений относительно оптимальных $\tilde{\mu}$ и $\tilde{\sigma}$:

$$\sum_{(i,j):\, b_{ij}=1} (a_{ij} - \tilde{\mu} - \tilde{\sigma}) + \sum_{(i,j):\, b_{ij}=0} (a_{ij} - \tilde{\mu}) = 0, \qquad \sum_{(i,j):\, b_{ij}=1} (a_{ij} - \tilde{\mu} - \tilde{\sigma}) = 0.$$

Подставляя второе уравнение в первое, находим, что $\tilde{\mu}$ равно *средней межкластерной связи*. Из второго уравнения вытекает, что $\tilde{\mu} + \tilde{\sigma}$ — это *средняя внутрикластерная связь*. При этом порог $\tilde{\rho}$ — их полусумма, $\tilde{\sigma}$ — характеристика контрастности связей (ср. с рис. 7).

Недостатком рассмотренного метода является наличие единого порога $\rho$ для всех классов, что не соответствует ситуации, когда присутствуют классы существенно различных размеров (о путях его преодоления рассказывается в [1, с. 175]).

## § 7. СРАВНЕНИЕ МЕТОДОВ

Определяющее значение при выборе метода классификации имеет общее число объектов $n$. Если оно велико (сотни или тысячи), то необходимо применять эвристический алгоритм А4 («Форель»), скомбинированный с А2 (кратчайший незамкнутый путь) (§ 2), или быстрые иерархические процедуры для редуктивных мер отдаленности (§ 4).

Для $n \leqslant 200$ в [52, с. 108–117] приводятся результаты экспериментального сравнения нескольких методов классификации

(в частности, А3 из § 2 и Р1—Р5 из § 3) путем их применения к моделированным совокупностям объектов.

Объекты генерировались как точки в $m$-мерном единичном гиперкубе ($3 \leqslant m \leqslant 7$). Задавалось число классов $k$ ($3 \leqslant k \leqslant 7$) и наудачу в гиперкубе выбирались центры классов $\boldsymbol{y}_l$ ($l = 1, \ldots, k$). Для каждого класса с помощью датчика случайных чисел моделировались $m$ длин сторон гиперпараллелепипеда с центром в $\boldsymbol{y}_l$, в который затем случайно бросались $n_l$ точек ($30 \leqslant n_l \leqslant 50$). При фиксированных $m$ и $k$ каждое разбиение генерировалось 50 раз и показатели усреднялись (всего было обработано около 2000 выборок).

Кроме того, генерировались «шумящие» объекты, равномерно распределенные в гиперкубе. Их количество составляло заданный процент $p$ от $n$ ($10\% \leqslant p \leqslant 30\%$). Эти объекты предназначались только для того, чтобы «сбивать с толку» алгоритмы классификации, но, естественно, не участвовали в расчете показателей качества классификации (одним из таких показателей было расстояние Хемминга (см. метрику D1 из § 1) между моделированным разбиением и разбиением, которое построил алгоритм).

Изложим кратко **основные выводы**. Наилучшей (почти идеальной) по восстанавливаемости разбиения проявила себя иерархическая процедура Р5 («метод Уорда»). Следом за ней идут процедура Р2 («дальнего соседа») и алгоритм А3 (метод $k$-средних Болла и Д. Холла) (случайный выбор эталонов в алгоритме А3 показал себя как крайне неудачный). Самой плóхой оказалась процедура Р1 («ближнего соседа»).

По уровню устойчивости к шуму лидерами стали алгоритмы А3 и Р5. Замыкает список снова Р1.

Не следует воспринимать эти результаты как приговор методу «ближнего соседа», а также близким к нему «по духу» (имеющим цепочечный эффект) алгоритмам А1 («связные компоненты») и А2 («кратчайший незамкнутый путь») из § 2. Речь может идти лишь о том, что на первичном этапе классификации, не обладая информацией о структуре классов, лучше применять менее чувствительные методы.

В случае, когда классы имеют сложную форму, скажем, относятся к типу С3 из § 1 («класс типа ленты или слабое сгущение»), именно алгоритмы Р1, А1 и А2 позволят правильно произвести разбиение.

Всякий необходимо причиняет пользу, употребленный на своем месте. Напротив того: упражнения лучшего танцмейстера в химии неуместны; советы опытного астронома в танцах глупы.

*Козьма Прутков*

Г. Миллиган и М. Купер в 1985 г. опубликовали результаты экспериментального сравнения в похожих условиях тридцати методов классификации (см. [52, с. 157]), в число которых вошли алгоритмы минимизации функционалов (см. § 5) $F4 = \left[\frac{1}{n-k}\,\mathrm{tr}\,\boldsymbol{W}_{int}\right] \Big/ \left[\frac{1}{k-1}\,\mathrm{tr}\,\boldsymbol{W}_{out}\right]$ (Калинский и Харабаш,

1974) и инвариантного функционала $F7 = \det \boldsymbol{W}_{int} / \det \boldsymbol{W}_{tot}$ (Фридман и Рубин, 1967). Первый оказался самым лучшим, а второй — в группе самых плохих. Дело, скорее всего, в удачной нормировке матриц рассеяния у $F4$ (сравните с формулой (9) гл. 16).

Вред или польза действия обусловливаются совокупностью обстоятельств.
*Козьма Прутков*

## § 8. ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ

После проведения классификации важно в удобной форме представить ее результаты. Приведем список важнейших (согласно [52, с. 159]) **характеристик классификации**.

1. Распределение номеров объектов по номерам классов.
2. Гистограмма межобъектных расстояний (подобная изображенной на рис. 7).
3. Средние внутриклассовые расстояния.
4. Матрица средних межклассовых расстояний.
5. Визуальное представление данных на плоскости двух (в пространстве трех) «наиболее информативных» признаков.
6. Дендрограмма для иерархических процедур.
7. Средние значения и размахи во всех классах для каждого признака.

Последний пункт наиболее принципиален, так как в подавляющем большинстве случаев интерпретация классов происходит по средним значениям признаков в них. Сопоставление же средних значений для заданного признака наиболее просто осуществляется, если классы не имеют наложения проекций. *Степень разделенности* классов по каждой оси можно охарактеризовать с помощью коэффициента

$$\gamma = 1 - \sum L_j / \sum R_l,$$

где $R_l$ — размах по заданному признаку $l$-го класса, а $L_1, L_2, \ldots$ — длины наложений проекций классов на ось признака (рис. 20). Если $\gamma = 1$, то классы полностью разделимы. Чем ближе $\gamma$ к 0, тем больше наложение проекций классов друг на друга.

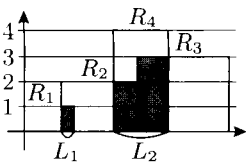


**Рис. 20**

## § 9. ПОИСК В ГЛУБИНУ

Основой для выделения связных компонент графа в методе А1 из § 2 служит алгоритм обхода всех вершин неориентированного связного графа $G$, именуемый *поиском в глубину* (сокращенно ПГ). Изложим его, следуя [28, с. 323].*)

В процессе поиска в глубину вершинам графа $G$ присваиваются номера (*ПГ-номера*), а ребра помечаются. В начале ребра не помечены, вершины не имеют ПГ-номеров. Начинаем с произвольной

*) В [28] приведены также подробные описания (с примерами) многих других полезных алгоритмов поиска на графах, скажем, *поиска в ширину* (с. 37) или *метода Дейкстры* нахождения кратчайшего пути между двумя заданными вершинами графа (с. 342).

## Задачи

1. Докажите, что метрика Чебышёва получается из метрики Минковского при $p \to \infty$.

2. Для выборки $X_1, \ldots, X_n$ из стандартного нормального распределения $N(0,1)$ запишите точно и вычислите приближённо константы, к которым сходятся по вероятности при $n \to \infty$ следующие последовательности:

а) $\alpha_n = MED\{X_1, \ldots, X_n\}$;

б) $\beta_n = MED\{|X_1|, \ldots, |X_n|\}$;

в) $\gamma_n = MED\{|X_1 - \alpha_n|, \ldots, |X_n - \alpha_n|\}$;

г) $\delta_n = X_{(3n/4)} - X_{(n/4)}$, где $X_{(1)} \le \ldots \le X_{(n)}$.

(В этой задаче нужны правильные ответы, а не строгие доказательства!)

3. Докажите, что минимум сумм расстояний от некоторой точки на плоскости до 4-х вершин квадрата достигается для точки пересечения диагоналей квадрата.

4. Добавьте к четырём вершинам квадрата 2 точки Штейнера так, чтобы построенный по 6 точкам КНП оказался короче, чем сумма длин двух диагоналей. (*Используйте оптимальное свойство точки Торричелли, из которой все стороны треугольника видны под углом* $120^o$.)

5. Докажите, что для меры отдалённости метода «дальнего соседа» выполняется неравенство треугольника.

6. Приведите пример, показывающий, что для меры отдалённости метода «ближнего соседа» неравенство треугольника может не выполняться.