Лекция 14

КЛАСТЕРНЫЙ АНАЛИЗ

Группировка исходных данных является основой решения задачи классификации и дальнейшей обработки информации. Традиционно из множества параметров, характеризующих объект, выбирается поначалу один, наиболее информативный с точки зрения исследователя, и производится группировка имеющихся данных по значениям этого параметра. Полученные группы — кластеры, по значениям второго по значимости параметра (опять-таки по мнению исследователя) разбиваются на подгруппы — более мелкие кластеры и т. д.

Если параметры ранжировать по степени важности затруднительно, применяется многомерная группировка, когда по нескольким исходным параметрам создается некоторый обобщенный признак (обобщенный параметр, обобщенный показатель), зависящий от исходных параметров. По этому обобщенному признаку осуществляется последующая классификация.

Возможно дальнейшее развитие классификации по нескольким обобщающим показателям, которые в этом случае носят название главных компонент. Соответствующий подход носит название *факторного анализа*.

Разделение множества образов на кластеры (*кластеризация*) — основа построения систем распознавания без учителя, причем выявление кластеров — «искусство весьма эмпирическое», так как в значительной степени качество алгоритма зависит от выбранной меры сходства образов и метода идентификации кластеров в системе признаков. *Кластер* (англ. *cluster*) – кисть, пучок, гроздь, куст, группа, рой, скопление. Это определение раскрывает основной смысл метода анализа: во всей совокупности данных нужно выделить группы — кластеры, которые как будто обладают сходными свойствами. Нахождение кластеров носит название кластер-анализа. Хотя принципиально методы кластерного анализа элементарны, их применение стало возможным только после появления современной вычислительной техники, так как эффективный поиск кластеров требует большого числа арифметических и логических операций.

Известны различные подходы к проблеме кластерного анализа, наиболее эффективным из которых является статистический подход

Мерами сходства являются расстояния, причем должны выполняться условия:

— расстояния между кластерами максимально возможные;

— расстояние между образами (объектами) внутри кластера минимально.

Задачи кластер-анализа можно разделить **на три группы** по исходным данным:

— число кластеров задано;

— число кластеров неизвестно и подлежит определению;

— число кластеров неизвестно и его определения не требуется, требуется лишь построение иерархического дерева, или дендрограммы.

Основной принцип кластеризации поясним на примере реализации ее наиболее простого алгоритма.

Пусть заданы N образов $\{\mathbf{x}_1,\ \mathbf{x}_2,\ ...,\ \mathbf{x}_N\}$. Такая запись соответствует N вектор-столбцам, отражающим набор значений параметров для одного объекта. Каждый из векторов $\mathbf{x}_i$ является набором из k значений параметров. Тогда весь набор данных может быть представлен в виде прямоугольной таблицы (матрицы) размером $k \times N$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kN} \end{pmatrix}.$$

Указанная форма представления данных не является единственной. Иногда исходную информацию задают в виде квадратной матрицы, например матрицы коэффициентов корреляции между отдельными параметрами. Ясно, что такая информация является производной от первоначальной.

Выберем в качестве центра первого кластера любой из образов, который условно обозначим через $\mathbf{x}_1 = \mathbf{z}_1$. Зададим порог распознавания C_0. Выберем из множества образов некоторый образ, который обозначим через $\mathbf{x}_2$. Вычислим расстояние между $\mathbf{x}_2$ и $\mathbf{z}_1$:

$$l_{21} = l(\mathbf{x}_2, \mathbf{z}_1) = \left\{ \sum_{j=1}^{M} \left| x_{2j} - z_{1j} \right|^{\nu} \right\}^{1/\nu}.$$

Если $l_{21} > C_0$, учреждается новый кластер с центром в точке $\mathbf{z}_2 = \mathbf{x}_2$, в противном случае образ $\mathbf{x}_2$ включается в кластер с центром в точке $\mathbf{z}_1$. Если $l_{21} > C_0$ и $\mathbf{z}_2$ — центр нового кластера, то для еще одного произвольно выбранного образа вычисляются расстояния l_{31} и l_{32} от $\mathbf{x}_3$ до $\mathbf{z}_1$ и $\mathbf{z}_2$ соответственно. Если $l_{31} > C_0$ и $l_{32} > C_0$, то учреждается новый

кластер с центром $\mathbf{z}_3 = \mathbf{x}_3$. В противном случае образ $\mathbf{x}_3$ зачисляется в тот кластер, чей центр к нему ближе. Те же операции повторяются с образами $\mathbf{x}_4$, $\mathbf{x}_5$ и т. д.

Эффективность алгоритма и результаты кластеризации существенно зависят от выбора центра первого кластера, порядка предъявления образов, значения порога C_0 и геометрических характеристик исходных данных.

Достоинствами приведенного алгоритма являются простое и быстрое получение оценок основных характеристик набора данных, однократный просмотр выборки для одного порога. Его недостатки: многочисленные эксперименты с разными значениями порога C_0 и различными исходными точками. Визуальная интерпретация, как правило, исключена. После каждого цикла вычислений целесообразен просмотр расстояний между центрами кластеров и количества образов в каждом кластере. Показателем успешного результата является выявление хорошо различимых кластеров.

Иерархические кластер-процедуры являются наиболее распространенными процедурами кластер-анализа. Эти процедуры разделяют на два класса:

— агломеративные;

— дивизимные.

В *агломеративных* процедурах начальным является разбиение на N одноэлементных классов, а конечным — разбиение из одного класса. На первом шаге каждое наблюдение $\mathbf{x}_i$ $(i = 1, 2, ..., N)$ рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы вычислительного алгоритма объединяются два самых близких кластера. После этого пересчитывается матрица расстояний, размерность которой на единицу меньше исходной. Такая последовательность действий происходит до тех пор, пока все кластеры не будут объединены в один класс.

В *дивизимных* процедурах последовательность разбиения обратная.

Последовательность объединений обычно графически характеризуется *дендрограммой*. Рассмотрим ее построение на примере. Пусть измерены значения двух параметров — обратного тока коллектора и коэффициента усиления в схеме с общим эмиттером у шести транзисторов. Данные отражены в таблице.

№ транзистора	1	2	3	4	5	6
$i_{к\ обр}$, мкА	5	6	5	10	11	10
h_{21}	10	12	13	9	9	7

Поскольку возможно графическое представление результатов на плоскости, целесообразно провести предварительный визуальный анализ данных. Не исключено, что кластеры будут видны без обработки данных.

Для решения задачи агломеративным иерархическим алгоритмом. Выберем в качестве меры расстояния между объектами евклидово расстояние.

Тогда расстояние между первым и вторым объектами составит

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24.$$

Вычисляя остальные расстояния и имея в виду, что $\rho_{ii} = 0$, можем построить матрицу расстояний

$$R_1 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}21 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 1 & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & 1 & 0 & 2{,}24 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}.$$

Из матрицы следует, что наименьшим расстоянием является $\rho_{45} = 1$, поэтому четвертый и пятый объекты объединяются в один кластер. Получим новую таблицу:

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4, 5)	(6)

Поскольку теперь кластер 4 является объединением объектов, необходимо договориться, как определить расстояние между кластерами в общем случае. Оно может быть определено по принципу ближайшего соседа, дальнего соседа, как расстояние между центрами, по принципу средней связи.

Воспользуемся расстоянием по принципу ближайшего соседа:

$$\rho_{\min}(S_l, S_m) = \min_{x_i \in S_l,\, x_j \in S_m} \rho(x_i, x_j).$$

В зависимости от специфики алгоритма это расстояние для составных кластеров может быть записано в разном виде. Наиболее простым является соотношение

$$\rho_{1,(4,5)} = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}\left|\rho_{14} - \rho_{15}\right| = 5{,}1.$$

Теперь матрица расстояний

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 2 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 0 \end{pmatrix}.$$

Теперь объединяем второй и третий объекты, имеем четыре кластера: $S_{(1)}$, $S_{(2,3)}$, $S_{(4,5)}$, $S_{(6)}$. Расстояние между кластерами вычисляем по той же формуле, например

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}\left|\rho_{(4,5),2} - \frac{1}{2}\rho_{(4,5),3}\right| = 5.$$

Продолжая вычисления получим следующее минимальное расстояние между кластерами (4, 5) и (6), равное 2, что приводит к новому кластеру (4, 5, 6) и новому наименьшему расстоянию между кластерами (1) и (2, 3), равному 2,24. Теперь получим два кластера (1, 2, 3) и (4, 5, 6), расстояние между которыми составит 5. Соответствующая дендрограмма будет иметь вид, показанный ниже.

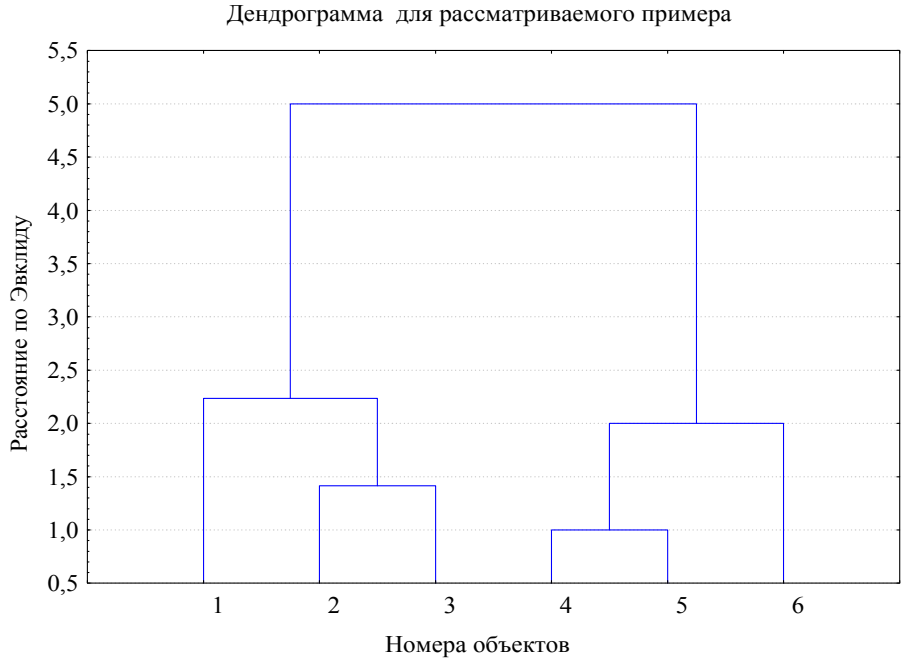

Рис. Л14.1. Дендрограмма

Задача Л14.1. 1. Последовательно рассчитать все необходимые расстояния, построить матрицы расстояний и провести последовательное объединение объектов в кластеры в соответствии с предложенными алгоритмами.

2. Найти приемлемый порог распознавания при объединении объектов в два кластера.

ВЫБОР ДИАГНОСТИЧЕСКИХ ПАРАМЕТРОВ (ПРИЗНАКОВ)

Определение статистической значимости диагностических параметров

Рациональный выбор диагностических параметров и формируемых на их основе признаков в значительной мере определяет успех диагностирования. Очевидно, что наиболее полезные параметры и признаки, которые инвариантны (нечувствительны) к изменениям внутри класса (кластера, диагноза) и резко меняются при переходе от класса к классу.

Может случиться так, что различные параметры (признаки) взаимосвязаны (коррелированны), так что их совместная регистрация не дает дополнительной полезной для постановки диагноза информации. Поскольку определение того или иного параметра (признака) требует определенных материальных и временных затрат, то на этапе разработки диагностической системы целесообразно выявить такие взаимосвязанные параметры и для диагностики бывает достаточно оставить один из такой группы параметров. Последнее несомненно удешевит разрабатываемую систему.

Установить степень коррелированности отдельных параметров x_k и x_l можно на основе определения коэффициентов по экспериментальным данным. Коэффициент корреляции R_{kl} между параметрами x_k и x_l определяется по формуле:

$$R_{kl} = \frac{\sum_{i=1}^{N}\left(x_{ki} - \overline{x}_k\right)\left(x_{li} - \overline{x}_l\right)}{\sqrt{\left\{\sum_{i=1}^{N}\left(x_{ki} - \overline{x}_k\right)^2\left(x_{li} - \overline{x}_l\right)^2\right\}}}$$

или

$$R_{kl} = \frac{\sum_{i=1}^{N}\left(x_{ki} - \overline{x}_k\right)\left(x_{li} - \overline{x}_l\right)}{(N-1)S(x_k)S(x_l)},$$

где N — число измерений этих параметров x_k и x_l при одних и тех же условиях; $\overline{x}_k$ и $\overline{x}_l$ — средние значения указанных параметров; $S(x_k)$ и $S(x_l)$ среднеквадратические отклонения x_k и x_l от своих средних значений.

Определив матрицу коэффициентов корреляции, строят *граф* (граф — система точек, соединенных отрезками; одна из простейших моделей взаимодействующих систем) корреляционных связей параметров, на основании которого определяют корреляционные *плеяды* параметров, значимо коррелирующих между собой. Из каждой плеяды выбирается один из параметров, который используют в дальнейшем для диагностирования. Выбор этого параметра определяется самим разработчиком, исходя из удобства и стоимости регистрации.

Для проверки статистической значимости коэффициента корреляции между параметрами x_k и x_l используют t-статистику:

$$t = \left|R_{kl}\right| \sqrt{\left(N-2\right) \Big/ \left(1-R_{kl}^2\right)}.$$

Если при заданном уровне значимости α вычисленная статистика превышает табличное значение $t_{\alpha,N-2}$ — $t > t_{\alpha,N-2}$, то коэффициент корреляции между этими параметрами признается статистически значимым с уровнем значимости α.

Построение графа корреляционных связей

Одним из возможных алгоритмов построения графа корреляционных связей включает следующие операции:

1. Находят наибольший по абсолютной величине недиагональный член корреляционной матрицы R_{ij}, $\left(i \neq j\right)$.

2. Формируют начало графа, задав точки, соответствующие x_i и x_j, соединив их прямой линией, над или под которой указывают значение R_{ij}.

3. В строках j и l находят следующий максимальный по абсолютной величине недиагональный член корреляционной матрицы R_{jl}, $\left(l \neq i, j\right)$.

4. Наносят точку, соответствующую x_l, и соединяют ее прямой с точкой x_j, указав над линией значение R_{jl}.

5. В строках i и j находят следующий максимальный по абсолютной величине недиагональный член корреляционной матрицы, кроме R_{ij} и R_{jl}, и т.д.

6. Продолжают указанную последовательность действий до тех пор пока число точек графа не станет равным N.

После построения графа определяют пороговое значение коэффициента корреляции $R_{\alpha,N-2}$ и разрывают связи с $R_{ij} < R_{\alpha,N-2}$. Последний определяют следующим образом. Задавшись уровнем значимости α, из статистических таблиц для t-статистики находят пороговое значение $t_{\alpha,N-2}$. Его подставляют в последнее соотношение

$$t_{\alpha,N-2} = \left|R_{\alpha,N-2}\right| \sqrt{\left(N-2\right) \Big/ \left(1-R^2_{\alpha,N-2}\right)}$$

и разрешают относительно коэффициента корреляции. Полученное значение является пороговым значением $R_{\alpha,N-2}$:

$$R_{\alpha,N-2} = \frac{t_{\alpha,N-2}}{t_{\alpha,N-2} + N - 2}.$$

Проиллюстрируем методику построения плеяд на примере следующей задачи.

Задача Л14.2. Электрический агрегат диагностируют с помощью восьми вибродатчиков, точки установки которых указаны на рис. Л14.2. Необходимо определить — все ли датчики необходимы для проведения диагностики.

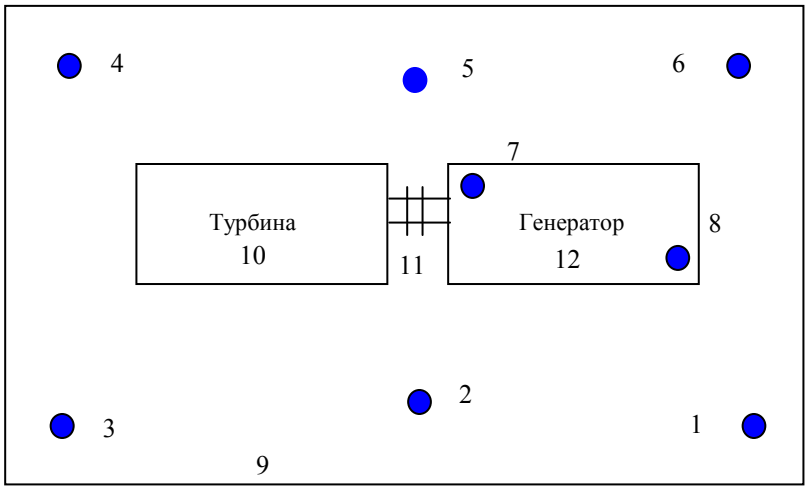

Рис. Л14.2. Расположение вибродатчиков при диагностике электрического агрегата

Матрица коэффициентов корреляции вибросигналов с датчиков приведена ниже.

Таблица Л14.1. Матрица коэффициентов корреляции

R_{kl}		**п а р а м е т р ы**							
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
п а р а м е т р ы	x_1	1,0	0,91	0,93	0,95	0,92	0,97	0,55	0,6
	x_2	0,91	1,0	0,88	0,89	0,95	0,91	0,38	0,37
	x_3	0,93	0,88	1,0	0,98	0,91	0,93	0,58	0,45
	x_4	0,95	0,89	0,98	1,0	0,91	0,94	0,44	0,62
	x_5	0,92	0,95	0,91	0,91	1,0	0,9	0,32	0,48
	x_6	0,97	0,91	0,93	0,94	0,9	1,0	0,66	0,54
	x_7	0,55	0,38	0,58	0,44	0,32	0,66	1,0	0,95
	x_8	0,6	0,37	0,45	0,62	0,48	0,54	0,95	1,0

Необходимо выявить корреляционные плеяды.

Находим максимальный недиагональный коэффициент корреляции $R_{\max} = R_{34} = 0,98$, $(i = 3,\ j = 4)$. Наносим точки x_3 и x_4, соединив их линией, над которой записываем коэффициент корреляции 0,98, см. рис. Л14.3. Далее, в строках 3 и 4 находим следующий наибольший коэффициент корреляции - $R_{41} = 0,95$. Наносим точку x_1, соединив ее с точкой x_4. Находим в строках 1 и 4 следующий наибольший коэффициент корреляции - $R_{16} = 0,97$, и указываем связь x_1 и x_6. В строках 1 и 6 находим $R_{15} = 0,92$, затем - $R_{52} = 0,95$, $R_{58} = 0,48$ и $R_{78} = 0,48$. В результате получаем граф корреляционных связей, показанный на рис. Л14.3.

Задав уровень значимости $\alpha = 0,05$, из статистических таблиц определяем пороговое значение t-статистики - $t_{0,05\ 6} = 2,45$. Пороговое значение коэффициента корреляции находим из уравнения

$$R^2_{0,05\,6}\frac{6}{1-R^2_{0,05\,6}} = t^2_{0,05\ 6} \simeq 6,0,$$

откуда $R_{0,05\,6} \approx 0,7$. Поэтому на графе корреляционных связей необходимо разорвать связи, у которых $R_{i\,j} < R_{0,05\,6}$. В результате образовались две корреляционные плеяды: $(x_1,...,x_6)$ и (x_7, x_8), выделенные на рис. Л14.3. Следовательно, для диагностики всего агрегата достаточно установки двух датчиков – одного из группы датчиков (1-6), другого – одного из дат-

чиков 7 или 8. Выбор этих датчиков определяется из удобства обслуживания и регистрации информации.

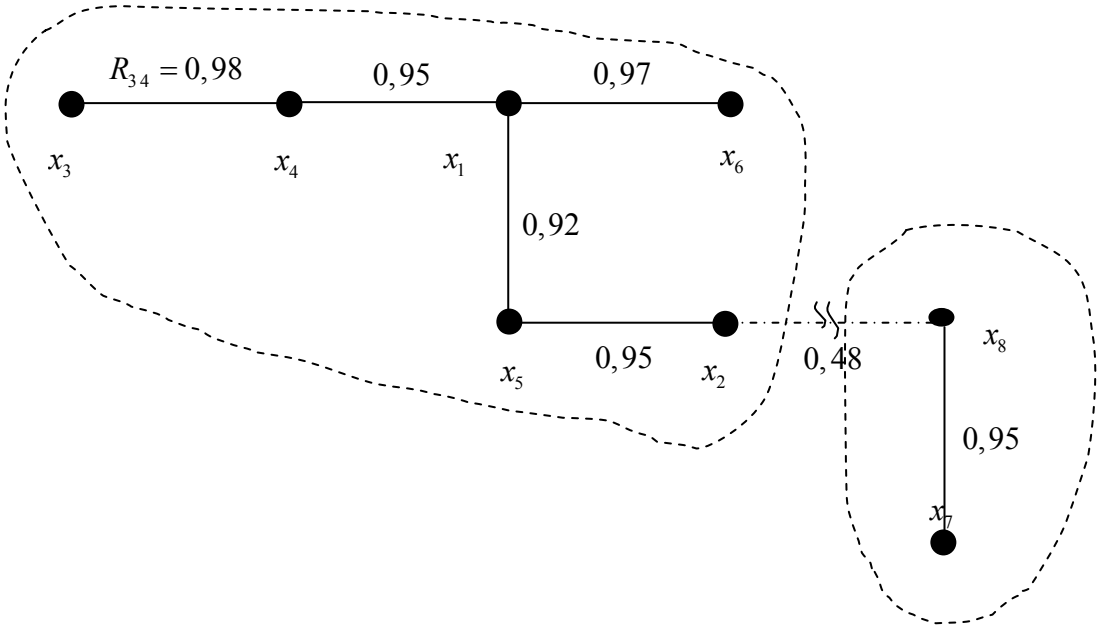

Рис. Л14.3.Граф корреляционных связей