Лекция 7

План лекции:

1. Особенности классификации объектов методом иерархического группирования.
2. Реализация метода иерархического группирования.

7.1 Метод иерархического группирования

Методы распознавания, где классы известны заранее и разделяющие функции вырабатывались в процессе обучения, сильно влияют на выбор признаков и критериев разделения, от которых зависит получаемый результат.

Для того чтобы уменьшить влияние первоначальных сведений, их обогащают дополнительной информацией. Например, уточняют пространственные или временные отношения (общепринятое пространственное отношение: глаза на лице находятся выше носа); находят существующие отношения между исследуемыми объектами (в частности, с помощью графов). Такие действия называются символическим описанием, которое получается в результате процедуры группирования, выполняющей функции и процедуры классификации.

Искомое символическое представление может иметь вид иерархической структуры, дерева минимальной длины или символического описания классов. Иерархия строится на основе понятия расстояния. Метод состоит в том, чтобы разработать последовательность разделений рассматриваемого множества на подгруппы, одна из которых обладает некоторым свойством, не присущим другим. Искомая иерархия основывается на предъявляемых выборках. Поскольку их число весьма велико, иногда на одном и том же множестве исходных данных могут быть получены различные иерархии. Известным примером служит иерархическая классификация в биологии по видам, родам, семействам, классам, типам, называемая «естественной» классификацией.

Рассмотрим правила построения иерархических группировок. Пусть X – множество, состоящее из m реализаций $\{X_1, X_2, ... X_m\}$, а $P(X)$ – множество всех его частей: $P(X) = \{0, X_1, \{X_1, X_2\}, \{X_1, X_3\}, ..., X_m\}$. Иерархией H называется подмножество, удовлетворяющее следующим условиям:

1. $X \in H$;
2. $\forall x_i \in X$, $x_i \in H$;
3. $\forall h, h' \in H$, если $h \cap h' \neq 0$, то либо $h \subset h'$, либо $h' \subset h$.

На рис. 1 приведен пример иерархии в виде дерева.

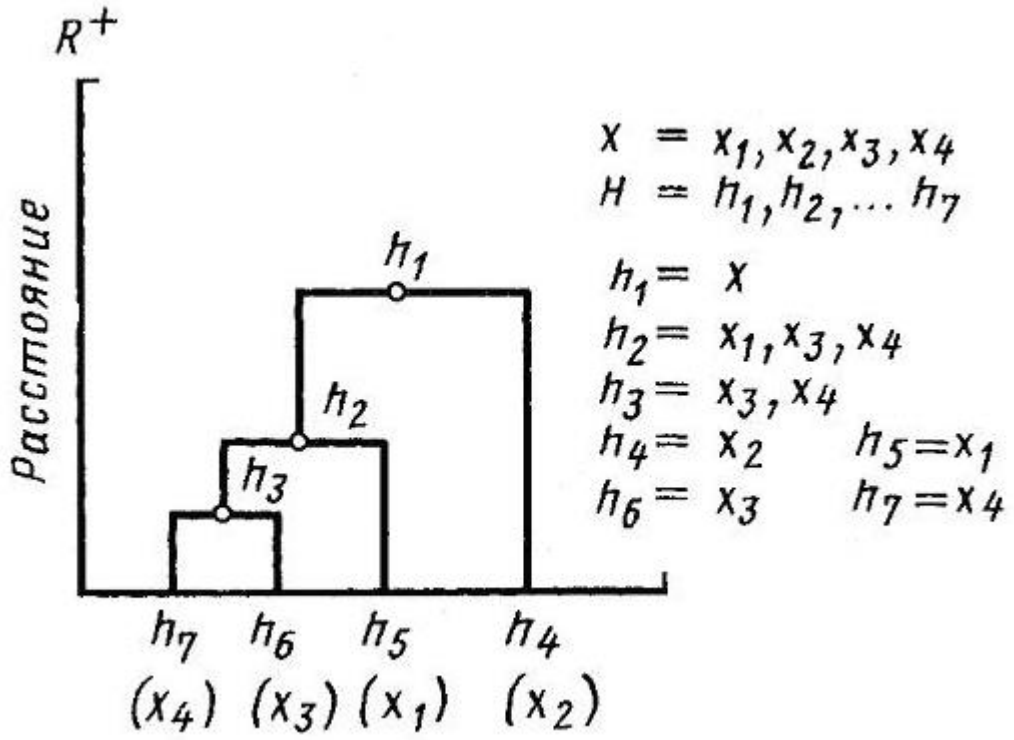

Рис. 1 – Иерархическое дерево

Здесь $h_1, h_2, \ldots h_7$ – элементы или вершины иерархического дерева; h_4, h_5, h_6, h_7 – терминальные элементы дерева H. Если терминальные элементы иерархии H содержат каждый только по одному элементу множества X, то они называются «атомами», а сама иерархия – «тонкой».

На практике чаще всего используется иерархия, обозначаемая вещественной функцией, откладываемой вдоль оси ординат. Эта функция называется расстоянием в широком смысле слова, поскольку она не связана с Евклидовым расстоянием между двумя точками. Выбор расстояния обусловливает построение иерархии.

Существует ряд алгоритмов для построения иерархических группировок и иерархий на их основе. Рассмотрим пример построения иерархии по критерию минимума. В этом случае иерархические группы A и B объединяются, если $d(A,D) = \inf\{d(A,p), d(B,q)\}$.

Даны четыре атома (x_1, x_2, x_3, x_4), расстояния между ними приведены в таблице 7.1.

Таблица 7.1

	x_1	x_2	x_3	x_4
x_1	0	5	0,5	2
x_2	5	0	1	0,6
x_3	0,5	1	0	2,5
x_4	2	0,6	2,5	0

На рис. 2 показано дерево, соответствующее исходным данным.

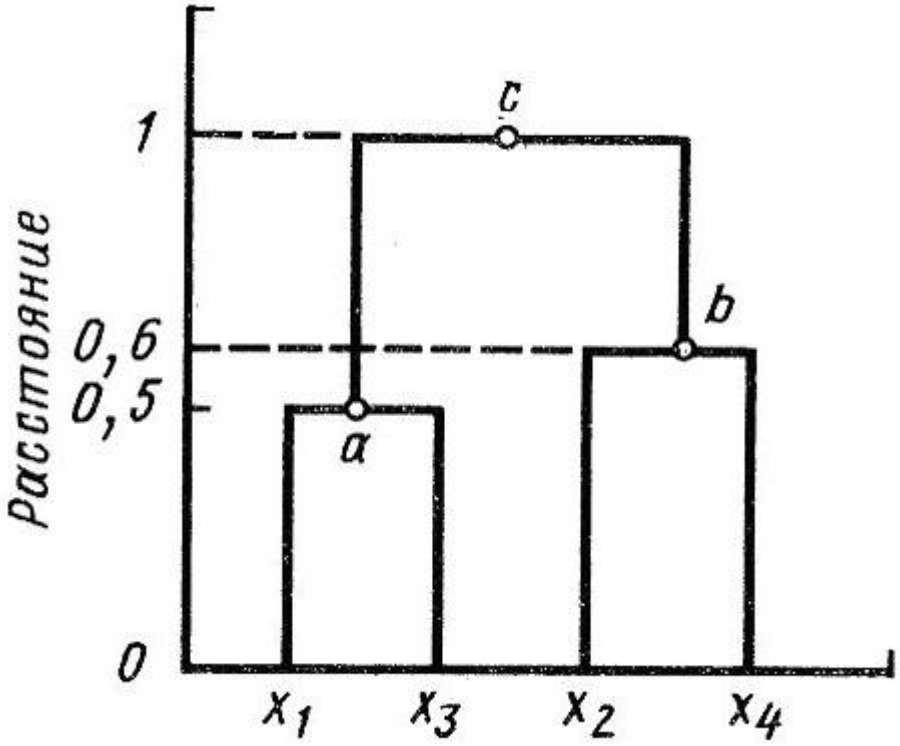

Рис. 2 – Результирующее иерархическое дерево

$d(x_1, x_3) = 0{,}5$ - минимальное расстояние, содержащиеся в таблице, следовательно, оно становится первым иерархическим объединением и обозначается $d(x_1, x_3) = \{a\}$, после чего элементы x_1 и x_3 в явном виде больше не участвуют в дальнейшем построении иерархии. Вместо них используется группировка a. Расстояния от нее до остальных элементов определяются следующим образом:

$$d\{a, x_2\} = \inf\{d(x_1, x_2),\ d(x_3, x_2)\} = 1;$$
$$d\{a, x_4\} = \inf\{d(x_1, x_4),\ d(x_3, x_4)\} = 2.$$

Продолжая процесс сокращения, выделяем новую группировку $(x_2, x_4) = b$, в результате остаются две группы a и b, объединяемые окончательно в $c = \{a, b\} = 1$.