机器学习

Machine learning

聚类分析

练习题答案

授课人：周晓飞
zhouxiaofei@iie.ac.cn

题目 1：数据（《机器学习》，周志华，2016， 表 9.1），层次聚类。

表 9.1 西瓜数据集 4.0

编号	密度	含糖率	编号	密度	含糖率	编号	密度	含糖率
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

层次聚类

层次聚类(hierarchical clustering)试图在不同层次对数据集进行划分，从而形成树形的聚类结构. 数据集的划分可采用“自底向上”的聚合策略，也可采用“自顶向下”的分拆策略.

AGNES 是一种采用自底向上聚合策略的层次聚类算法. 它先将数据集中的每个样本看作一个初始聚类簇，然后在算法运行的每一步中找出距离最近的

两个聚类簇进行合并, 该过程不断重复, 直至达到预设的聚类簇个数. 这里的关键是如何计算聚类簇之间的距离. 实际上, 每个簇是一个样本集合, 因此, 只需采用关于集合的某种距离即可. 例如, 给定聚类簇 C_i 与 C_j, 可通过下面的式子来计算距离:

$$\text{最小距离: } d_{\min}(C_i, C_j) = \min_{\boldsymbol{x} \in C_i, \boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) , \tag{9.41}$$

$$\text{最大距离: } d_{\max}(C_i, C_j) = \max_{\boldsymbol{x} \in C_i, \boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) , \tag{9.42}$$

$$\text{平均距离: } d_{\mathrm{avg}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\boldsymbol{x} \in C_i} \sum_{\boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) . \tag{9.43}$$

显然, 最小距离由两个簇的最近样本决定, 最大距离由两个簇的最远样本决定, 而平均距离则由两个簇的所有样本共同决定. 当聚类簇距离由 $d_{\min}$、$d_{\max}$ 或 d_{avg} 计算时, AGNES 算法被相应地称为“单链接”(single-linkage)、“全链接”(complete-linkage) 或“均链接”(average-linkage)算法.

AGNES 算法描述如图 9.11 所示. 在第 1–9 行, 算法先对仅含一个样本的初始聚类簇和相应的距离矩阵进行初始化; 然后在第 11–23 行, AGNES 不断合并距离最近的聚类簇, 并对合并得到的聚类簇的距离矩阵进行更新; 上述过程不断重复, 直至达到预设的聚类簇数.

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇距离度量函数 d;
聚类簇数 k.

通常使用 $d_{\min}$, $d_{\max}$ 或 d_{avg}.

过程:

初始化单样本聚类簇.

1: **for** $j = 1, 2, \ldots, m$ **do**
2: $\quad C_j = \{\boldsymbol{x}_j\}$
3: **end for**

初始化聚类簇距离矩阵.

4: **for** $i = 1, 2, \ldots, m$ **do**
5: $\quad$ **for** $j = 1, 2, \ldots, m$ **do**
6: $\quad\quad M(i, j) = d(C_i, C_j)$;
7: $\quad\quad M(j, i) = M(i, j)$
8: $\quad$ **end for**
9: **end for**
10: 设置当前聚类簇个数: $q = m$
11: **while** $q > k$ **do**

$i^* < j^*$.

找出距离最近的两个聚类簇 C_{i^*} 和 C_{j^*};
合并 C_{i^*} 和 C_{j^*}: $C_{i^*} = C_{i^*} \bigcup C_{j^*}$;
for $j = j^* + 1, j^* + 2, \ldots, q$ **do**
将聚类簇 C_j 重编号为 C_{j-1}
end for
删除距离矩阵 M 的第 j^* 行与第 j^* 列;
for $j = 1, 2, \ldots, q-1$ **do**
$M(i^*, j) = d(C_{i^*}, C_j)$;
$M(j, i^*) = M(i^*, j)$
end for
$q = q - 1$
end while

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

图 9.11 AGNES 算法

以西瓜数据集 4.0 为例, 令 AGNES 算法一直执行到所有样本出现在同一个簇中, 即 $k = 1$, 则可得到图 9.12 所示的"树状图" (dendrogram), 其中每层链接一组聚类簇.

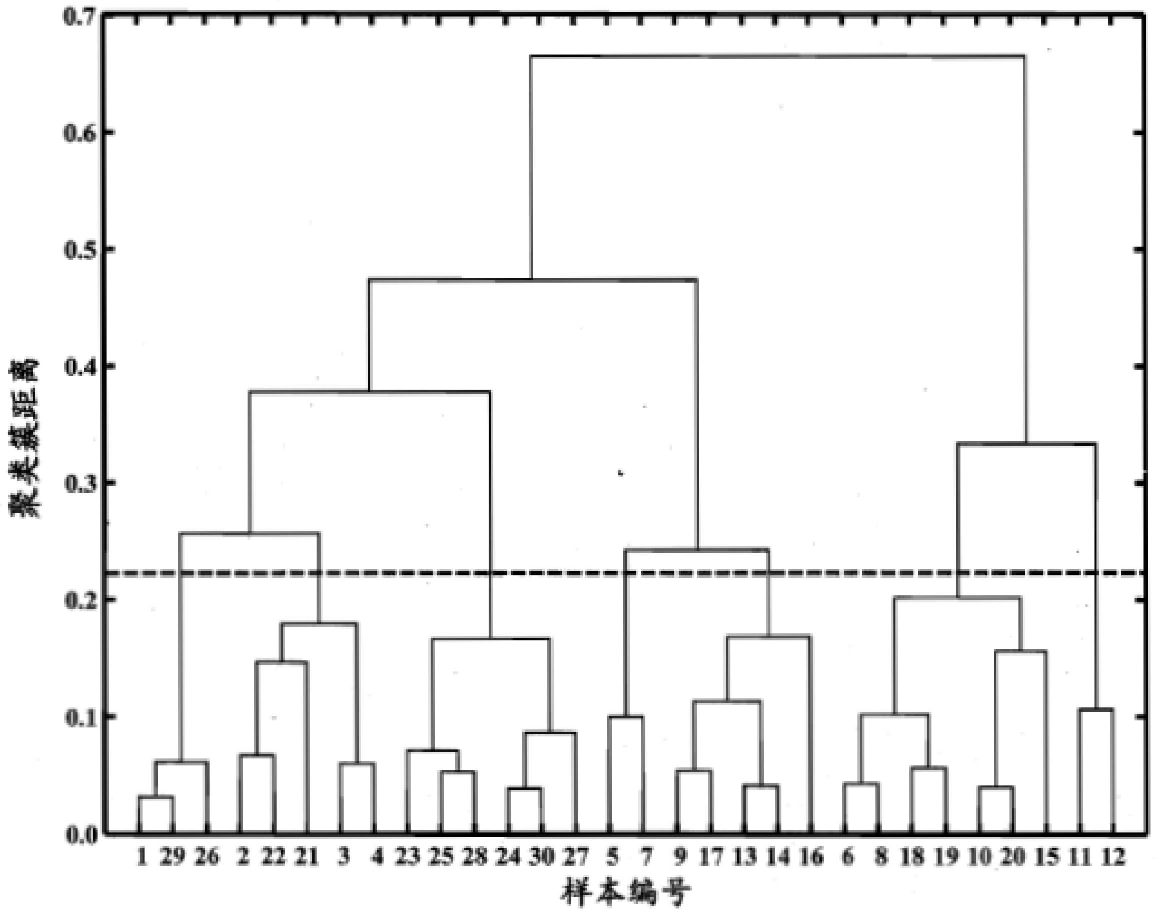


图 9.12 西瓜数据集 4.0 上 AGNES 算法生成的树状图(采用 $d_{\max}$). 横轴对应于样本编号, 纵轴对应于聚类簇距离.

在树状图的特定层次上进行分割，则可得到相应的簇划分结果. 例如，以图 9.12 中所示虚线分割树状图，将得到包含 7 个聚类簇的结果:

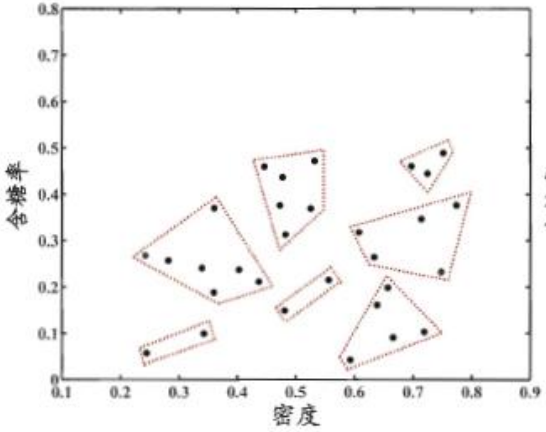
$$C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_{26}, \boldsymbol{x}_{29}\};\ C_2 = \{\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_{21}, \boldsymbol{x}_{22}\};$$

$$C_3 = \{\boldsymbol{x}_{23}, \boldsymbol{x}_{24}, \boldsymbol{x}_{25}, \boldsymbol{x}_{27}, \boldsymbol{x}_{28}, \boldsymbol{x}_{30}\};\ C_4 = \{\boldsymbol{x}_5, \boldsymbol{x}_7\};$$

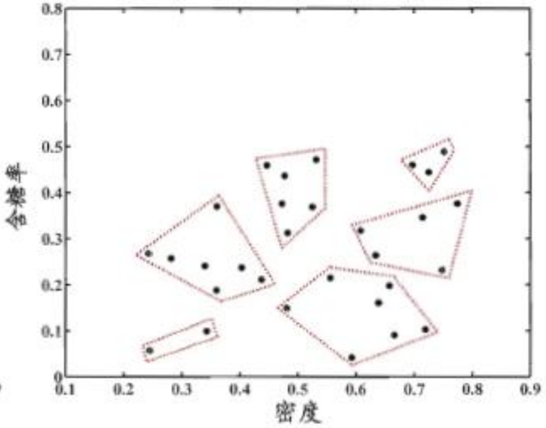
$$C_5 = \{\boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}, \boldsymbol{x}_{16}, \boldsymbol{x}_{17}\};\ C_6 = \{\boldsymbol{x}_6, \boldsymbol{x}_8, \boldsymbol{x}_{10}, \boldsymbol{x}_{15}, \boldsymbol{x}_{18}, \boldsymbol{x}_{19}, \boldsymbol{x}_{20}\};$$

$$C_7 = \{\boldsymbol{x}_{11}, \boldsymbol{x}_{12}\}.$$

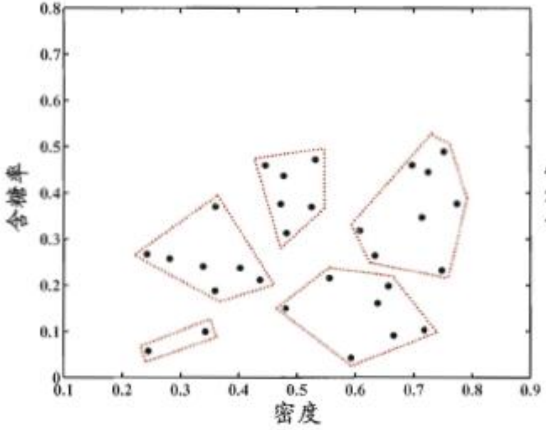
将分割层逐步提升，则可得到聚类簇逐渐减少的聚类结果. 例如图 9.13 显示出了从图 9.12 中产生 7 至 4 个聚类簇的划分结果.



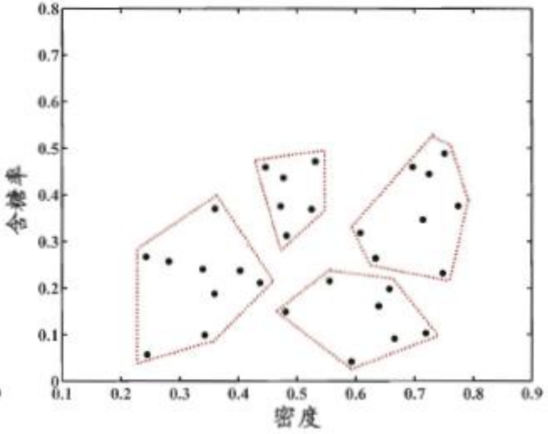
(a) 聚类簇数 $k = 7$



(b) 聚类簇数 $k = 6$



(c) 聚类簇数 $k = 5$



(d) 聚类簇数 $k = 4$

图 9.13 西瓜数据集 4.0 上 AGNES 算法(采用 $d_{\max}$)在不同聚类簇数($k = 7, 6, 5, 4$)时的簇划分结果. 样本点用“●”表示, 红色虚线显示出簇划分.

题目 2：使用 k-means 算法，给出下列数据的聚类结果。

点	x_1	x_2
P1	0	1
P2	1	2
P3	2	2
P4	8	8
P5	9	10
P6	10	10

注：初始化聚类中心为 P1 和 P2。

解：

第一轮：

{P1}，{P2}

{P1}，{P2，P3}

{P1}，{P2，P3，P4}

{P1}，{P2，P3，P4，P5}

{P1}，{P2，P3，P4，P5，P6}

新的质心：(0, 1), (6, 6.4)

第二轮：

{P1}，{}

{P1, P2}, {}

{P1, P2, P3}, {}

{P1, P2, P3}, {P4}

{P1, P2, P3}, {P4, P5}

{P1, P2, P3}, {P4, P5, P6}

新的质心：(1，5/3)，(9，28/3)

第三轮：

{P1}，{}

{P1，P2}，{}

{P1，P2，P3}，{}

{P1，P2，P3}，{P4}

{P1，P2，P3}，{P4，P5}

{P1，P2，P3}，{P4，P5，P6}

新的质心：（1，5/3），（9，28/3）

质心不再改变，得出最终的聚类结果：

{P1，P2，P3}，{P4，P5，P6}

题目 3：使用自底向上层次聚类，给出下列数据的聚类结果，簇之间的相似度采用簇质心的距离。

点	x_1	x_2
P1	0	1
P2	1	2
P3	2	2
P4	8	8
P5	9	10
P6	10	10

解：

开始每一个点为一类：

{P1}，{P2}，{P3}，{P4}，{P5}，{P6}

对应的聚类质心坐标为：

（0,1），（1,2），（2,2），（8,8），（9,10），（10,10）

经过计算，{P2}与{P3}之间的距离最小，进行合并：

{P1}，{P2，P3}，{P4}，{P5}，{P6}

对应的聚类质心坐标为：

（0,1），（3/2,2），（8,8），（9,10），（10,10）

经过计算{P5}，{P6}之间的距离最小，进行合并：

{P1}，{P2，P3}，{P4}，{P5，P6}

对应的聚类质心坐标为：

（0,1），（3/2,2），（8,8），（19/2,10）

经过计算{P1}，{P2，P3}之间的距离最小，进行合并：

{P1，{P2，P3}}，{P4}，{P5，P6}

对应的聚类质心坐标为：

（1,5/3），（8,8），（19/2,10）

经过计算{P4}，{P5，P6}之间的距离最小，进行合并：

{P1，{P2，P3}}，{P4，{P5，P6}}

最后两个集合进行合并，得到最终的聚类结果：

{{P1，{P2，P3}}，{P4，{P5，P6}}}