
电子科技大学

# 电子科技大学（深圳）高等研究院

# 模式识别

| | |
|---|---|
| 专业 | 控制科学与工程 |
| 班级 | 5 班 |
| 学生 | 陈玉熙 202122280534 |
| 学生 | 李育泓 202122280515 |
| 教师 | 凡时财 |
| 日期 | 2021.1.5 |

# 目　　录

# 1 实验要求

1. 采用 C 均值聚类算法对男女生样本数据中的身高、体重、50m 成绩 3 个特征进行聚类分析，考察不同的类别初始值以及类别数对聚类结果的影响，并以友好的方式图示化结果。

2. 采用分级聚类算法对男女生样本数据进行聚类分析。尝试采用身高，体重、50m 成绩 3 个特征进行聚类，并以友好的方式图示化结果。

原始数据（部分）图如图 3-1 所示：

| 编号 | 性别 男1女0 | 籍贯 | 身高(cm) | 体重(kg) | 鞋码 | 50米成绩 | 肺活量 | 喜欢颜色 | 喜欢运动 | 喜欢文学 | 喜欢数学 | 喜欢模式识别 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 湖北 | 163 | 51 | 41 | 7.5 | 2500 | 蓝 | 1 | 1 | | |
| 2 | 1 | 河南 | 171 | 64 | 41 | 7.5 | 3500 | 蓝 | 0 | 0 | | |
| 3 | 1 | 云南 | 182 | 68 | 45 | 7.8 | 4900 | 蓝 | 1 | 0 | | |
| 4 | 1 | 广西 | 172 | 66 | 42 | 8.2 | 4800 | 绿 | 0 | 1 | | |
| 5 | 1 | 四川 | 185 | 80 | 44 | 8.5 | 5100 | 蓝 | 0 | 0 | | |
| 6 | 0 | 河北 | 164 | 47 | 38 | 9 | 2500 | 紫 | 1 | 1 | | |
| 7 | 0 | 河南 | 160 | 46 | 38 | 9 | 2500 | 白 | 1 | 1 | | |
| 8 | 1 | 重庆 | 170 | 46 | 41 | 7 | 3000 | 蓝 | 1 | 1 | | |
| 9 | 1 | 重亲 | 178 | 60 | 41 | 7 | 4200 | 绿 | 0 | 0 | | |
| 10 | 1 | 江苏 | 180 | 71 | 43 | 7.5 | 3500 | 紫 | 0 | 0 | | |

图 3-1 原始数据（部分）图

# 2 数据处理

使用 Python 语言编程，调用 xlrd 库的函数来读取 excel 文件中对应工作表中的数据，将身高、体重、50m 成绩 3 个特征的数据存储在数组中。

由于原始数据存在缺失值或异常值等异常数据，所以需要先对数据进行清洗。同时由于原始数据中各数据相对独立，不存在关键数据，所以并未考虑采取插补法、建模法等办法来修复数据，而是直接将异常数据删除。

# 3 C 均值聚类

首先，分别计算身高、体重、50m 成绩的均值以及标准差，对数据进行归一化处理。

其次，编写主体逻辑程序实现 C 均值聚类算法。先选取 C 个点作为初始中

心点，代表 C 类不同的点，循环计算所有原始数据点与这 C 个中心点的欧式距离，将每个点归入与其最近的中心点所属的类。遍历完所有数据点之后，重新计算每个类的新中心点，再进行上述的操作，直到某一次循环中心点不再发生变化，算法结束。

最后，将每个样本的三维特征映射为空间中的点，画出三维空间中的散点图，将不同类别的数据点用不同颜色标注，每个类的中心点用红色三角形符号代表。同时，由于归一化后的数据比较抽象，无法直观体现数据本身的实际意义，在绘图时将数据重新变换为原来的尺度。C 均值聚类部分效果图如图 3-2 所示。

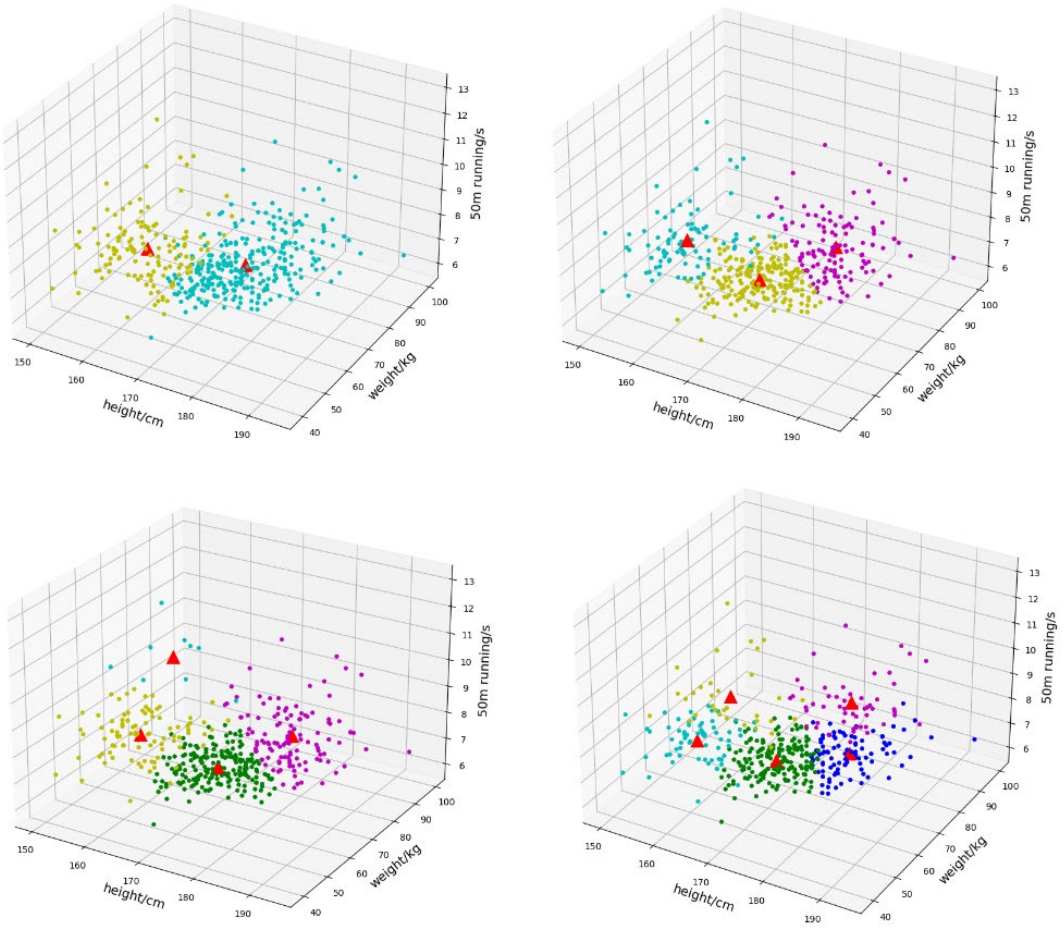



图 3-2 C 均值聚类部分效果图

成功实现 C 均值聚类基本功能后，对以下几个方面进行了进一步分析：

（1）分类准确率。虽然聚类属于无监督学习，但在这个实验中，其实数据是带有标签的。所以在 C 取值为 2，即分为两类时，可以使用数据标签来计算分类准确率。最终发现分为两类时，分类准确率为 88.81%，类别中心点如表 3-1 所

示。

表 3-1 分类中心点表（C=2）

| 类别 | 身高 | 体重 | 50m 成绩 |
|---|---|---|---|
| 1 | 164.2 | 52.6 | 8.6 |
| 2 | 175.4 | 68.8 | 7.4 |

其中值得注意的是数据的真实标签可能与分类预测的标签相同，也可能相反，需要进行判断。

（2）归一化。由于原始数据三维特征尺度不一，理论上直接进行聚类效果会比归一化后的效果差。依旧取 C=2，用真实标签进行准确率计算进行实验，实际也证明确实如此。归一化前后聚类结果表如表 3-2 所示。

表 3-2 归一化前后聚类结果表（C=2）

| | 类别 | 身高 | 体重 | 50m 成绩 | 准确率 |
|---|---|---|---|---|---|
| 归一化前 | 1 | 167.81 | 56.0 | 8.0 | 63.70% |
| | 2 | 177.3 | 73.9 | 7.6 | |
| 归一化后 | 1 | 164.2 | 52.6 | 8.6 | 88.81% |
| | 2 | 175.4 | 68.8 | 7.4 | |

（3）类别初始值。在类别数 C 固定的情况下，选取不同类别初始值时，虽然中心点最后都会一样，但是迭代次数会不一样。

实验最开始选取中心点的逻辑是，求出每个特征最大值与最小值的差，这个差值乘以一个随机数，再加上最小值作为中心点。这种取法是每个特征维度单独取值，实际上可能导致某个点取在整个数据点的范围之外。在类别数 C 选择较大时，可能会出现所有数据点都归不到某个中心点，即某一类为空的情况。

为避免上诉情况，初始中心点的选取可以规定在已有数据点中任意选取，实验证明如此选择不会出现某一类为空的情况。

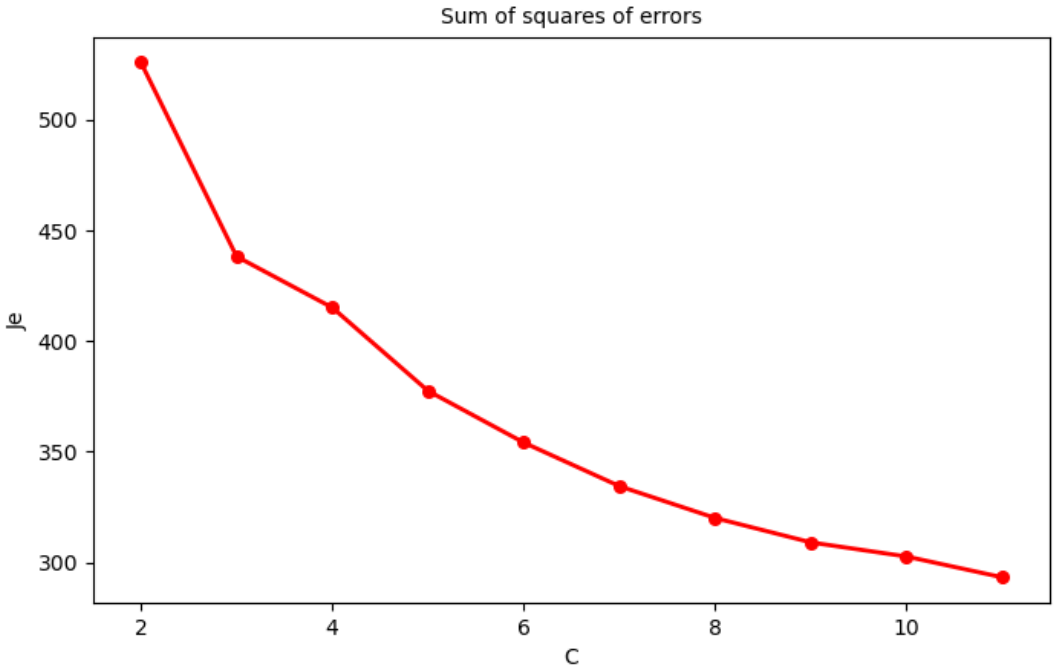
但是随机选取时很多时候迭代次数均比较大，实验最终实现的初始中心点选取方法如下。首先随机选择一个点作为第一个初始类中心点，然后选择距离该点最远的那个点作为第二个初始类中心点，然后再选择距离前两个点的最近距离中最大的点作为第三个初始类的中心点，以此类推，直至选出 C 个初始类中心点。

在实验多此尝试中，发现这种方法比随机选取平均迭代次数更少。

（4）C 的选择。C 均值聚类为无监督的算法，C=2 时使用标签计算准确率进行评价是本实验的特殊情况，在一般情况下，可以使用基于误差平方和准则 Je 对类内距离进行评价。对 C 连续取值进行聚类操作，分别计算每一类中的数据点与中心点的距离之和，再进行相加得到 Je 值。Je 计算具体公式如下：

$$J_e = \sum_{i=1}^{c} \sum_{k=1}^{N} \|y_k - m_i\|^2$$

不同 C 值下的 Je 曲线图如图 3-3 所示。




理论上 C 的取值足够大，大到每一个点都自成一类，Je 便为 0，所以 Je 的值也并不是越小越好，一般认为在 C 值不断增加过程中，Je 从迅速变小到减慢的拐点处，接近为最优 C 值，实验认为 C=3 为近似最优值。

# 4 分级聚类

分级聚类，也称层次聚类，是一个通用的聚类算法系列，通过依次合并或拆分自身来构建嵌套的聚类。这种聚类层次结构表示为树（或树状图）。树的根是

收集所有样本的唯一簇，叶子是只有一个样本的簇。聚类过程中逐级考查类间相似度，依此决定类别数。

对于逐级合并的方法，一般先把所有样本各自视作一类，然后逐级聚合成一类。这种方法被称为“聚合法”（Agglomerative）。

对于逐级分解的办法，一般先把所有样本视作一类，逐级分解为每一样本一类。这种方法被称为“分解法”（Divisive）。

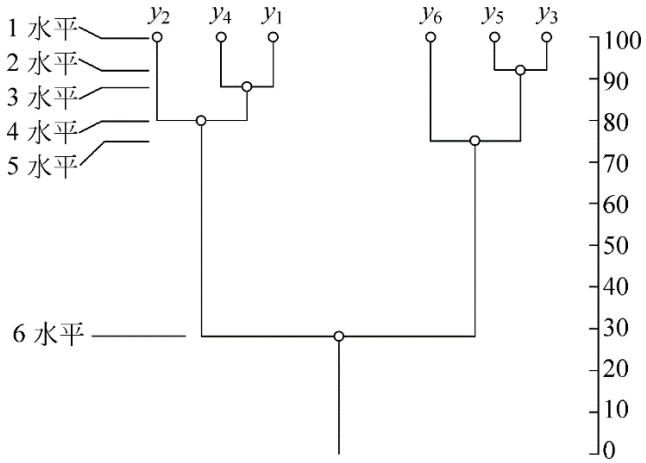



图 4.1 分级聚类结构图

通常来说，我们使用聚合法进行聚类，其算法流程如下：

（1）初始化，每个样本形成一类

（2）把相似性最大（距离最小）的两类合并

（3）重复（2），直到所有样本合并为两类

而在分级聚类算法中，很关键的一点就是类间相似性度量，它取决了分级聚类的质量，一般常用的几种度量方式如下：

1. 最近距离（single） $\Delta(\Gamma_i, \Gamma_j) = \min\limits_{\substack{y\in\Gamma_i \\ \tilde{y}\in\Gamma_j}} \delta(y, \tilde{y})$

2. 最远距离（complete） $\Delta(\Gamma_i, \Gamma_j) = \max\limits_{\substack{y\in\Gamma_i \\ \tilde{y}\in\Gamma_j}} \delta(y, \tilde{y})$

3. 均值距离（average） $\Delta(\Gamma_i, \Gamma_j) = \delta(m_i, m_j)$

4. 加权距离（weighted） $d(u, v) = (\mathrm{dist}(s, v) + \mathrm{dist}(t, v))/2$

5. 质心距离（centroid） $\mathrm{dist}(s, t) = \|c_s - c_t\|_2$

6. 类内方差最小化（ward）

$$d(u,v)=\sqrt{\frac{|v|+|s|}{T}d(v,s)^2+\frac{|v|+|t|}{T}d(v,t)^2-\frac{|v|}{T}d(s,t)^2}$$

不同相似性对度量的结果往往影响很大，例如最近距离和最远距离对比结果：

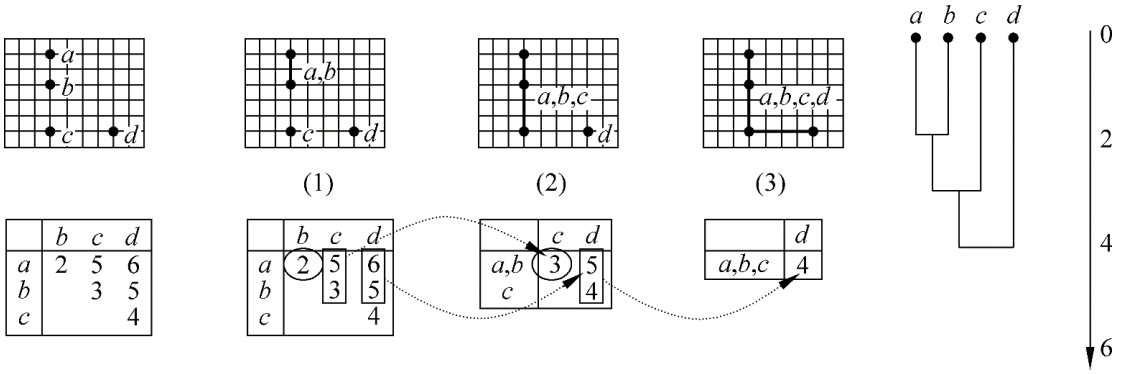



图 4.2 最近距离连接

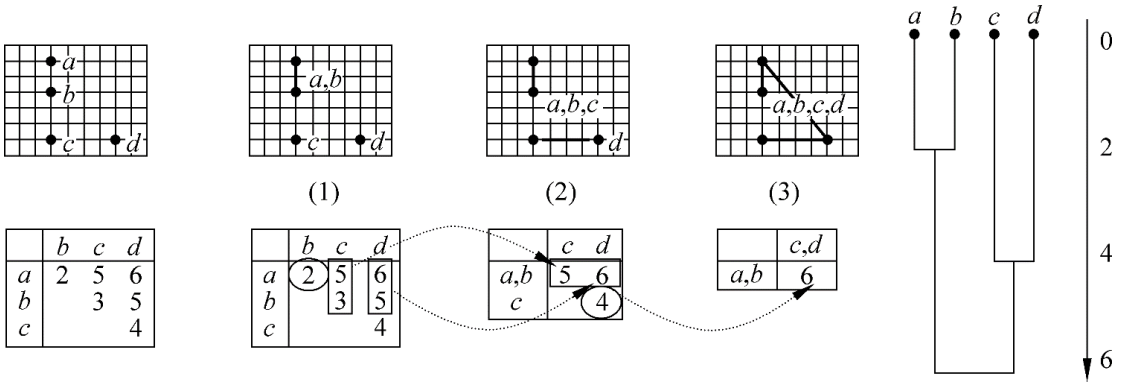



图 4.3 最远距离连接

当然，除了欧氏距离，还有其它度量方式，这里就不再讨论了。

另外，值得注意的是：分级聚类是一种局部搜索算法，对样本中的噪声较为敏感。而且，聚类树的画法不是唯一的。同一类中的两个分支可以左右互换而不改变聚类结果，但会改变树的外观和分析者的判断。

接下来介绍我们的实验：实验要求进行采用身高，体重、50m 成绩 3 个特征进行聚类，我们首先将三个特征数据从表格导出，和前两次实验一样，进行数据预处理和特征工程。然后正式开始实验。

实验第一步是选择聚类算法，这里我们使用的是 Agglomerative 聚合法，自底向上逐级合并，度量为欧氏距离。具体度量方式为 ward（即最小化被合并的集群的方差）。初始距离阈值为 0，即从完全分散分类开始逐级合并。

通过调整合并层次，我们可以分别得到聚类结果如下：

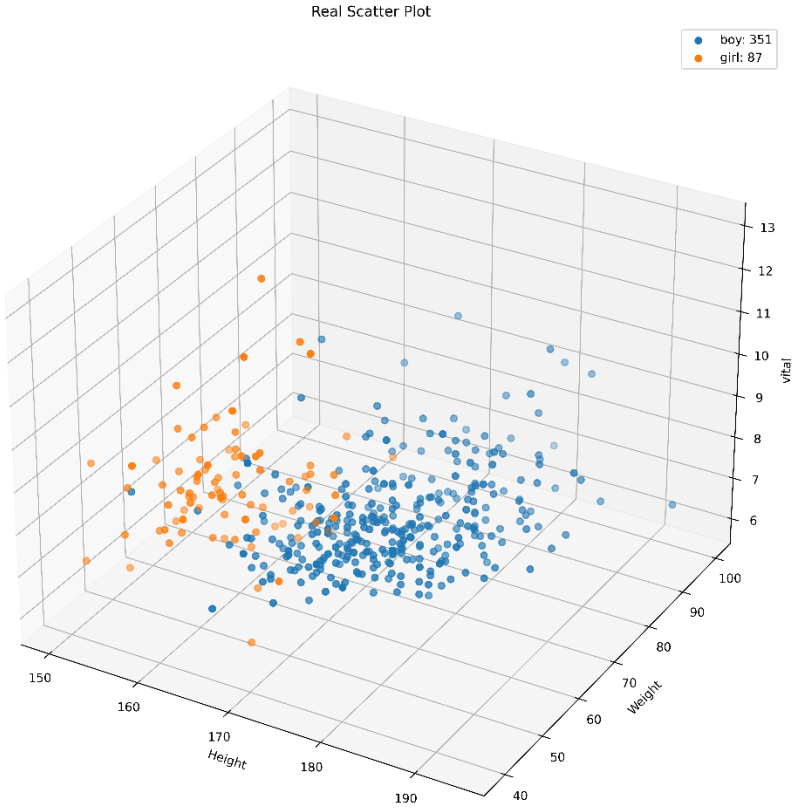



图 4.4 真实标签

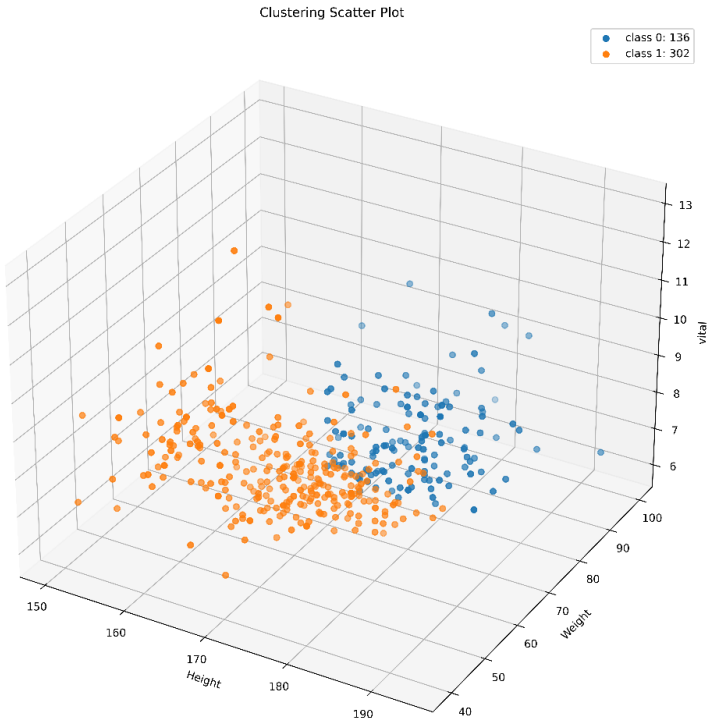



图 4.5 零级合并（树状图 n-1 级）

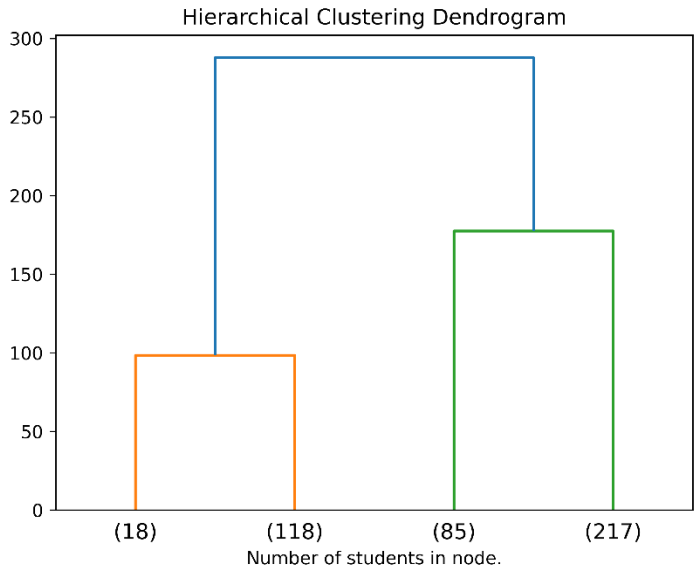



图 4.6 二级合并（树状图 n-3 级）

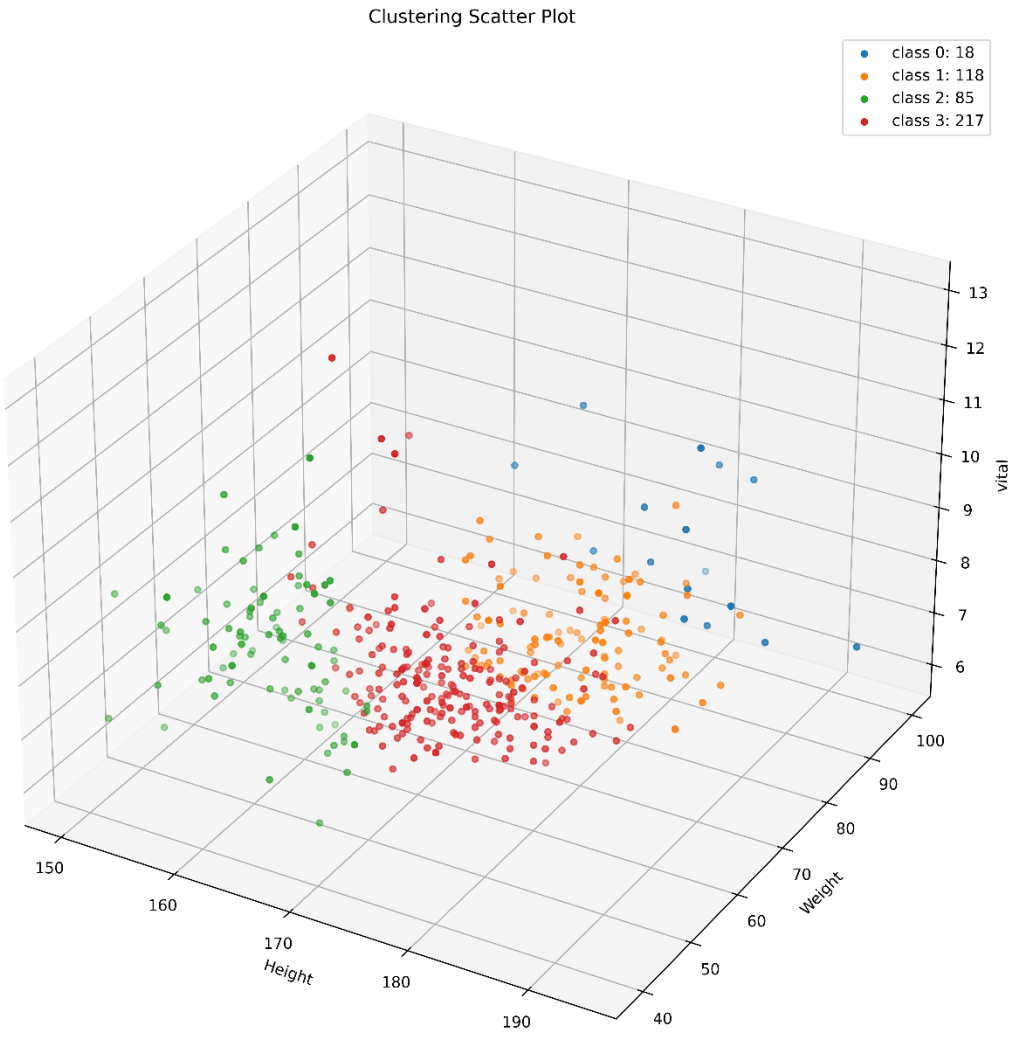



图 4.7 二级合并（树状图 n-3 级）

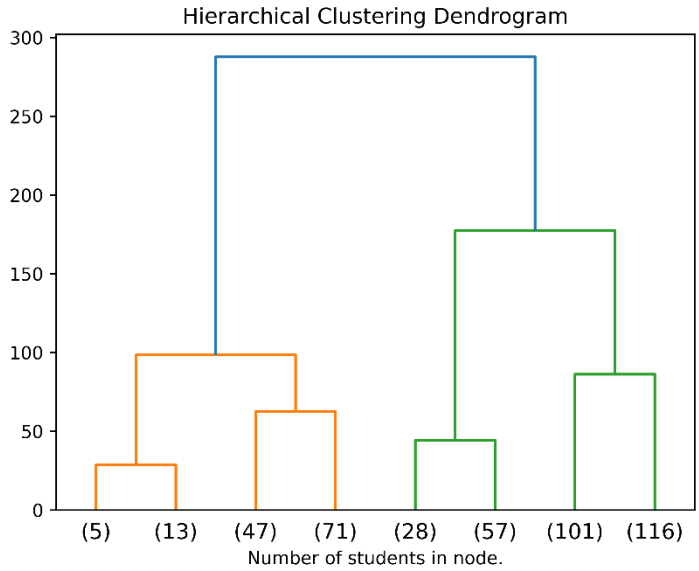



图 4.8 六级合并（树状图 n-7 级）

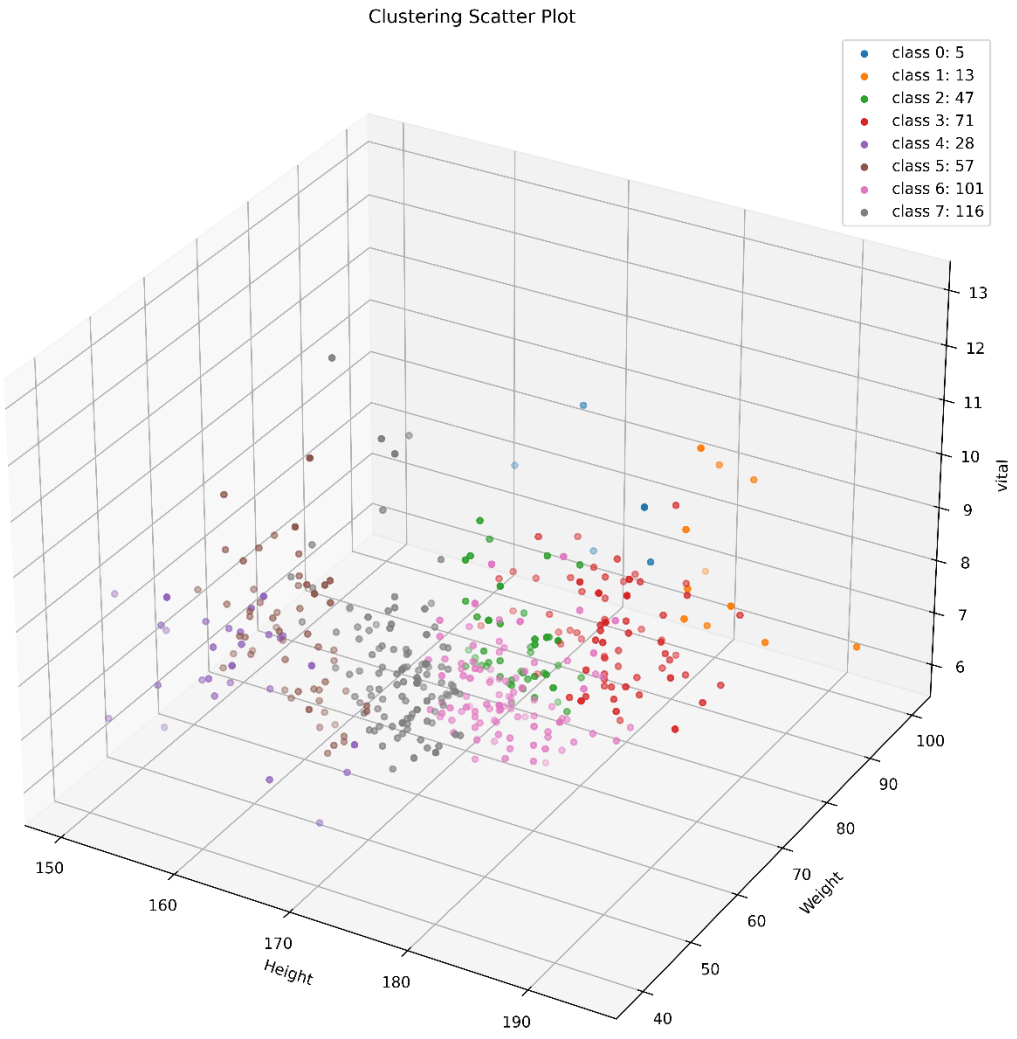



图 4.9 六级合并（树状图 n-7 级）

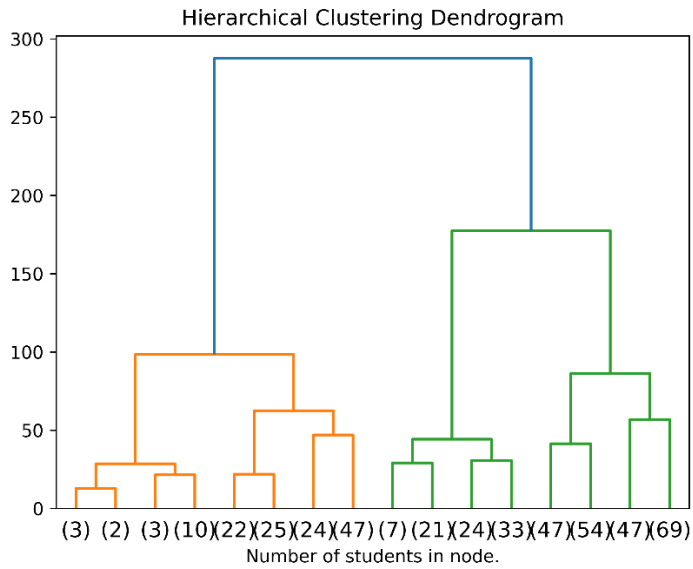



图 4.10 十四级合并（树状图 n-15 级）

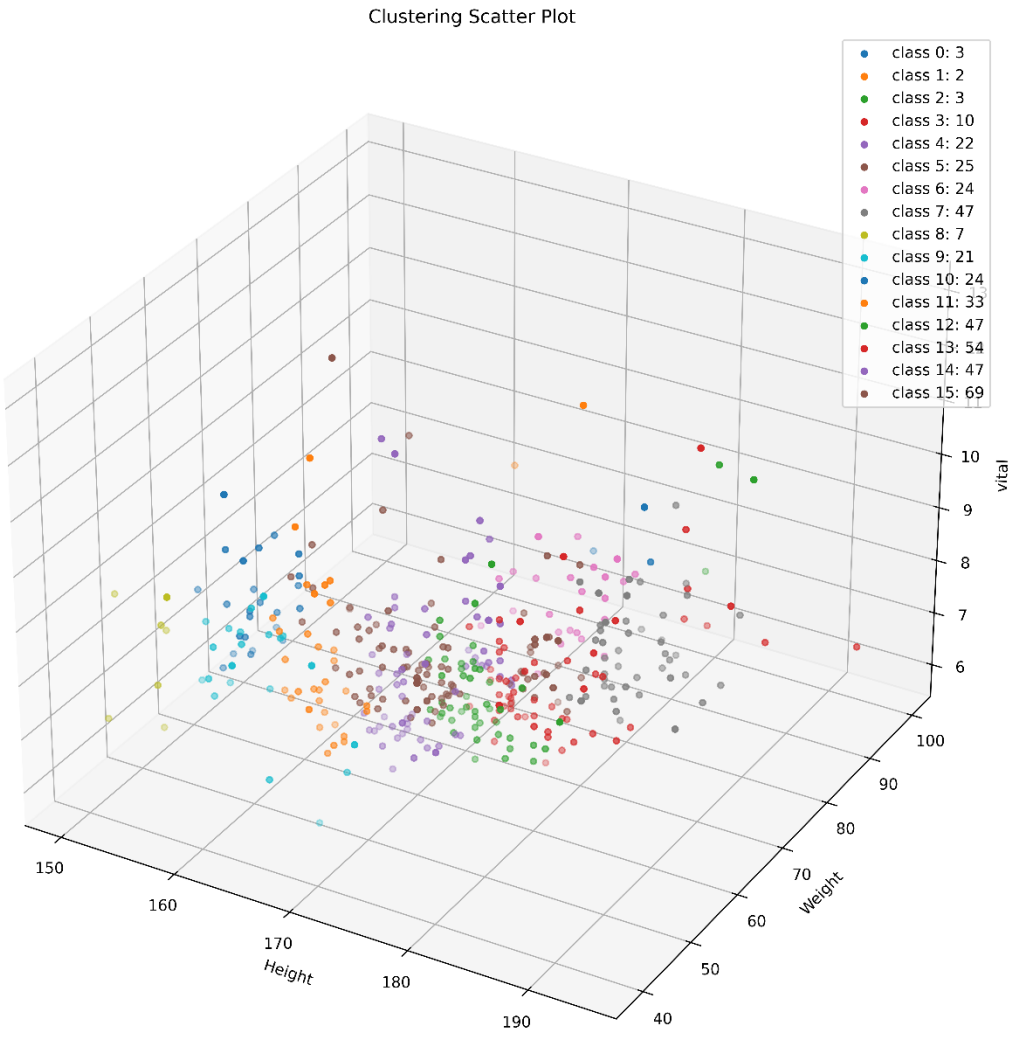



图 4.11 十四级合并（树状图 n-15 级）

按照课上所说，我们聚合过程停止的标准实际上有一下三种：

a. 若事先已知类型数目 $c$，则应停留在 n-c+1 级

b. 若预先给定类间距离阈值 $D_0$，当类间距离大于阈值，则终止聚类过程

c. 若 $c$、$D_0$ 均不知，先完成 n→1 全部聚类过程，然后根据类间相似性等，获取适当的聚类级作为最佳结果。

这里因为我们题设分类标准为男女性别，应该按照 a 标准作为聚类终止标准。聚类最终结果应该选取树状图 n-1 级（图 4.6 蓝色与橙色绿色交点）。以此获得的分类散点图如图 4.5 所示。

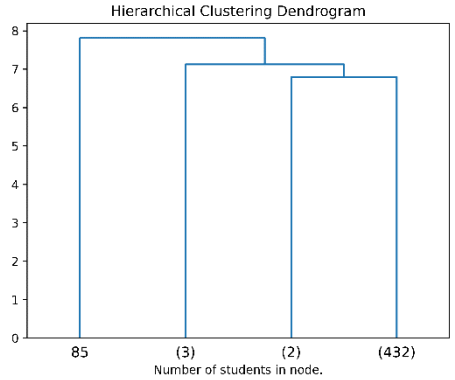
我们将结果与图 4.4 的原始标记对比，很容易看出，单纯使用此三种特征进行分级聚类，选择 n-1 级的二分类结果与原始性别标记差别较大。

分析原因，一开始我们认为是不是分级聚类局部搜素的偶然性导致可能存在不确定性的结果输出。后来经过多次实验和进一步的理论认识，明白在样本数据和度量标准一定的情况下，这种搜索算法的的结果是一定的。

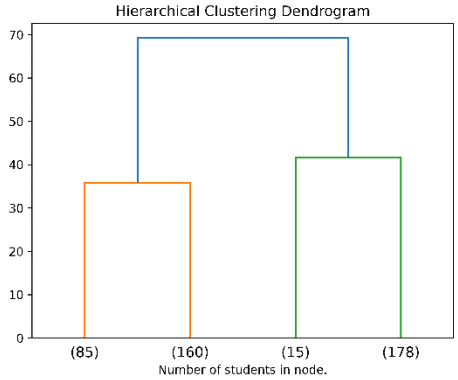
进一步思考，我们将>=n-3 级结果树状图（4.6）和散点图（4.7）绘制出并与之对比，发现了错误的主要原因。由于递归部分代码和时间的限制，我们只能按照等效完全二元树的结构进行绘图。但是还是很容易从 4.6 看出，如果是 n-2 级三分类的时候，令 class0、class1 的集合为 class5。计算 class2、class3、class5 之间的类间距，认为 class2、class3 间距最小，对二者进行合并。而实际上 class2 包含了绝大多数女生样本，在这一步应该 class3 与 class5 合并才是正确的。

我们又进一步检索，分别得到>=n-7 级和>=n-15 级的树状图及散点图结果（图 4.8-图 4.11），更加深入的理解分级聚类的搜索算法过程。可以发现除了 n-2 级的分类错误，其它的聚类过程基本都是符合实际的，这也验证了之前所说的：作为局部搜索算法，分级聚类对样本中的噪声较为敏感。一点小小的扰动、一步的搜索错误就导致了最后输出的较大差距。

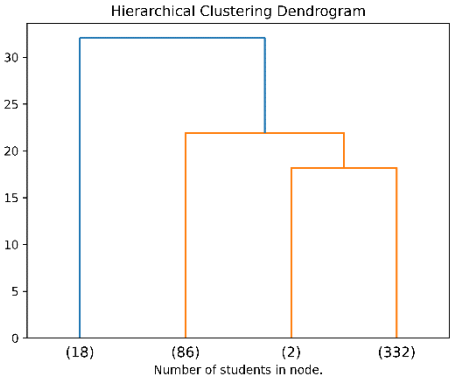
针对这样的错误，我们尝试不再使用 warp 度量，我们分别尝试了上述度量方式，为了更容易看出问题，我们选择>=n-3 级分类结果显示，最终结果如下：



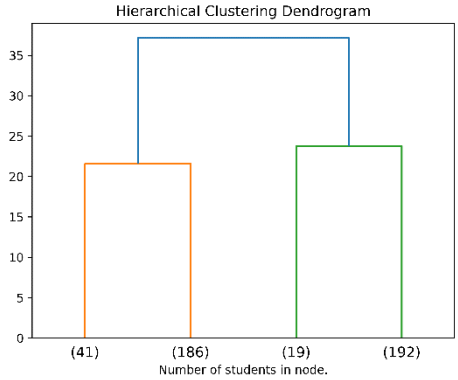

最近距离（single）



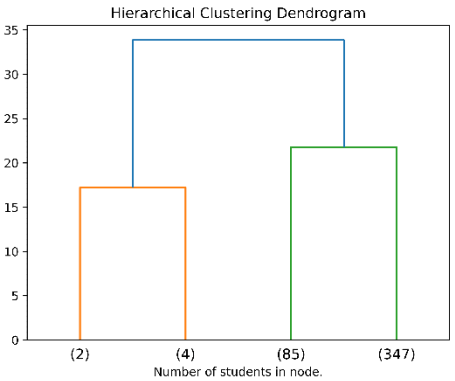

最远距离（complete）



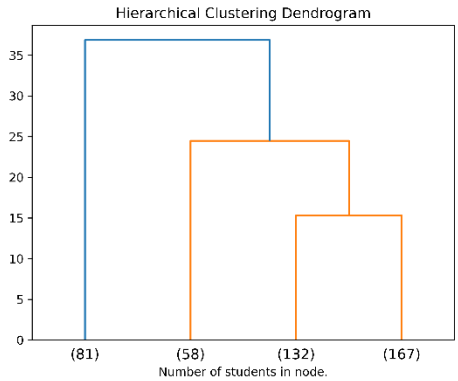

均值距离（average）




加权距离（weighted）




质心距离（centroid）

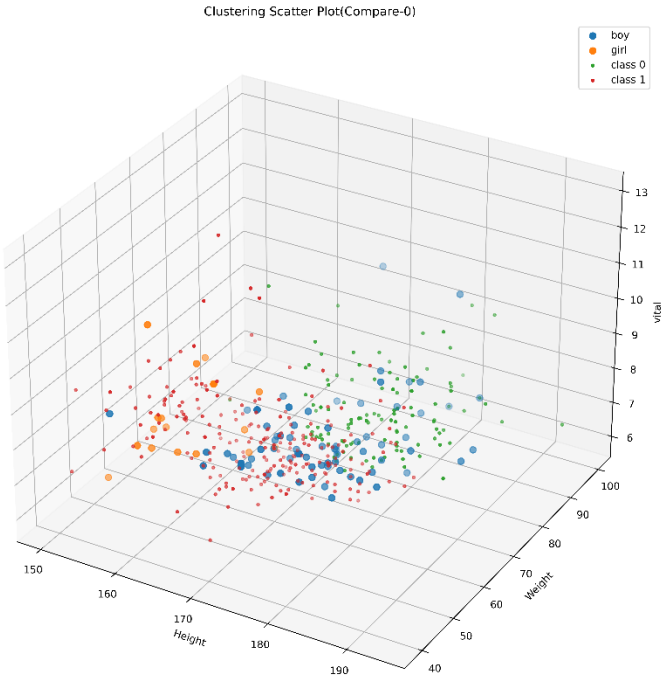


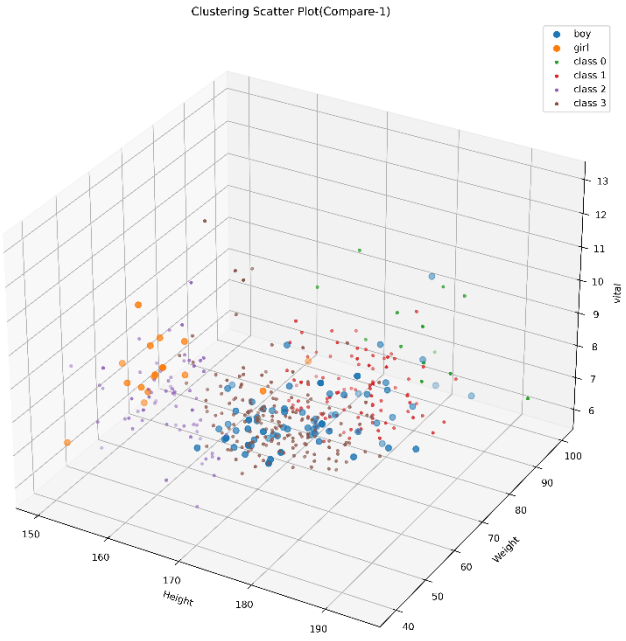

中位数距离（centroid）

可以看出，在最远距离（n-3 级）、均值距离（n-2 级）、质心距离（n-2 级）、中位数距离（n-1 级）时，均完成了绝大多数样本的正确分类。但是只有中位数距离真正完成了 n-1 级（二分类）阶段的多数正确分类。其它算法都会因为某一

级的错分导致最后结果的较大差距。这也说明分级聚类进行前一定要进行足够的数据预处理，保证输入样本尽量没有离群点以及其它类型的噪声影响分类结果。

最后，我们以随机 8：2 对样本抽样，以 20%样本点对 80%样本分级聚类的大概区域范围进行测试，测试结果图如下（由于 3D 渲染时间太长，没有用色域标识区域，仅以点群大致判断）：




分级聚类结果与实际样本对比

# 5 分工

陈玉熙：负责分级聚类的实现与报告撰写。

李育泓：负责 C 均值聚类的实现与报告撰写。