

阈值点，并将该阈值点作为分割阈值点。连续特征的决策树分裂指标如表 9-4 所示。

计算分割阈值点的具体步骤如下：

- (1) 确定可能的阈值集合。
 - (a) 对特征的取值按升序进行排序。
 - (b) 将两个特征取值之间的中点作为可能的分割点。（如果有 n 条训练样本，那么可能有 $n - 1$ 个阈值，过多的阈值会对应大量的指标计算。我们可以通过判断两个相邻特征取值所对应的样本的分类信息是否相同来减少计算次数。如果该分类信息相同，则该点不适合作为分割点，可以略过，这样能显著减少阈值的数量。）
- (2) 通过计算分类指标，确定指标最优的阈值，并将其作为分割阈值点。
 - (a) 对于阈值 a_j ，则 $(-\infty, a_j]$ 和 $(a_j, +\infty)$ 就是该特征的划分子集，计算分类指标。
 - (b) 找出指标最优的阈值，并将其作为分割阈值点。

表 9-4 连续特征的决策树分裂指标

算法	分裂指标
ID3	使用信息增益 (Information Gain, IG) $IG(C, A^{(i)}) = H(C) - H(C A^{(i)})$ 选择使 $IG(C, A^{(i)})$ 最大的特征属性 $A^{(i)}$ 以及阈值分割点 $a^{(i)}$
C4.5	使用信息增益率 (Information Gain Ratio, IGR) $IGR(C, A^{(i)}) = \frac{IG(C, A^{(i)})}{H(A^{(i)})}$ 选择使 $IGR(C, A^{(i)})$ 最大的特征属性 $A^{(i)}$ 以及阈值分割点 $a^{(i)}$
CART	使用基尼指数 (Gini Index) $Gini(C, A^{(i)})$ 考虑基尼增量： $\Delta Gini(C) = Gini(C) - Gini(C, A^{(i)} = a_j^{(i)})$ 目标是使基尼增量最大，即选择基尼指数 $Gini(C, A^{(i)} = a_j^{(i)})$ 最小的属性特征 $A^{(i)}$ 以及阈值分割点 $a^{(i)}$

9.2.3 指标计算示例

我们以 Golf 数据集（见表 9-1）为例，目标是根据某一天的天气状态（天气的整体情况、