决策树算法总结

虹 *

2023 年 4 月 29 日

本文概要

决策树是一种十分常见的分类和回归算法，决策树学习是从训练集中归纳出一种分类规则，并以树状图形的方式表现出来，可以看成为一个 if-then 规则的集合．决策树分类通常包含 ID3、C4.5、CART 三种算法，三者选择最优特征的方法有所不同：其中 ID3 算法使用的是信息增益；C4.5 算法使用的是信息增益比；CART 算法使用的是基尼指数．三种算法适用的场景也各不相同：

- ID3 算法：适用于离散型数据．ID3 算法使用信息增益来选择最佳的分裂变量，因此需要将数据集离散化为有限的分类．因此，ID3 算法通常用于文本分类、垃圾邮件分类等离散型数据的分类问题．
- C4.5 算法：适用于离散型和连续型数据．与 ID3 算法不同，C4.5 算法使用信息增益比来选择最佳的分裂变量，能够处理连续型和离散型的数据，因此适用范围更广．C4.5 算法常用于数据挖掘、信用评分等领域．
- CART 算法：适用于离散型和连续型数据．与 ID3 算法不同，C4.5 算法使用信息增益比来选择最佳的分裂变量，能够处理连续型和离散型的数据，因此适用范围更广．C4.5 算法常用于数据挖掘、信用评分等领域．

关键词：ID3　C4.5　CART　决策树减枝　递归算法

*微信公众号：数学与智能科学

1 三种算法的优缺点

1.1 三种算法的优点

- ID3 算法：算法简单易懂，计算效率高．对于数据缺失的情况有很好的处理能力．
- C4.5 算法：
  - 支持离散型和连续型数据的处理．
  - 采用信息增益比来选择最佳分裂变量，能够避免 ID3 算法中选择取值较多的属性作为分裂变量的问题．
  - 采用剪枝操作，能够有效地防止过拟合．
- CART 算法：
  - 能够处理离散型和连续型数据的分类和回归问题．
  - 采用基尼指数来选择最佳分裂变量，能够更好地处理连续型数据
  - 采用剪枝操作，能够有效地防止过拟合．

1.1.1 三种算法的缺点

- ID3 算法：
  - 对于连续型数据和缺失数据的处理能力不足．
  - 容易产生过拟合，不能很好地应对噪声数据．
- C4.5 算法：
  - 对于噪声数据的处理能力不足．
  - 计算效率较低，需要对数据进行多次扫描．
- CART 算法：
  - CART 算法生成的是二叉树，对于多分类问题需要进行二次划分，增加了计算复杂度．
  - 对于缺失数据的处理能力有限．

2 算法框架的使用

Algorithm 1 ID3 算法

Input: 训练数据集 D，特征集 A.

Initialize:

some text goes here ...

1: **while** *not converged* **do**

2: ...

3: **end while**

4: **Output:** 决策树 T.

3 数学定理定义的使用

定理 **1.** 设 a, b 是两个实数，则 $2ab \leqslant a^2 + b^2$

证明. 因为 $(a-b)^2 \geqslant 0$
所以可得到 $a^2 + b^2 - 2ab \geqslant 0$, 从而得到 $2ab \leqslant a^2 + b^2$. □

引理 **1.** 引理 *1*

命题 **1.** 命题一