Data Mining

Model Overfitting

Introduction to Data Mining, 2nd Edition

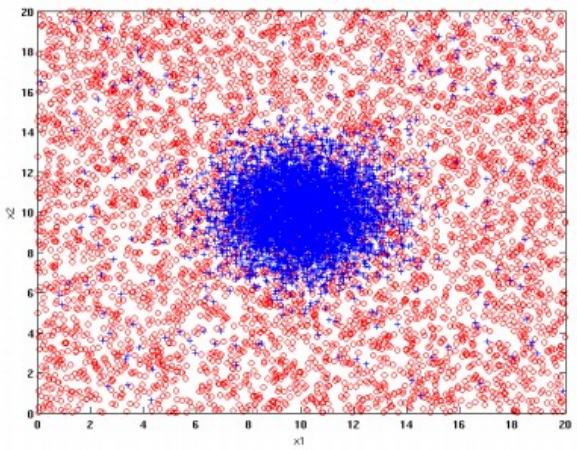
by

Tan, Steinbach, Karpatne, Kumar

1

02/05/2020 数据挖掘导论，第 2 版 2
分类错误
训练错误(明显错误)训练集中发生的错误
测试错误
测试集上提交的错误
泛化误差
从同一分布中随机选择记录时模型的预期误差

Example Data Set



Two class problem:

+ : 5400 instances

- 5000 instances generated from a Gaussian centered at (10,10)
- 400 noisy instances added

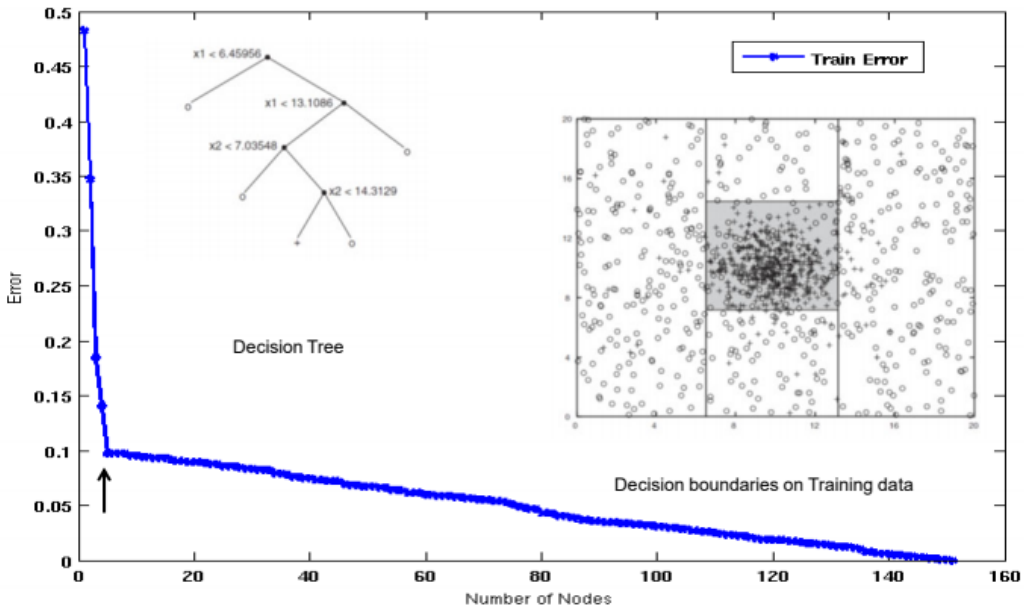
o : 5400 instances

- Generated from a uniform distribution

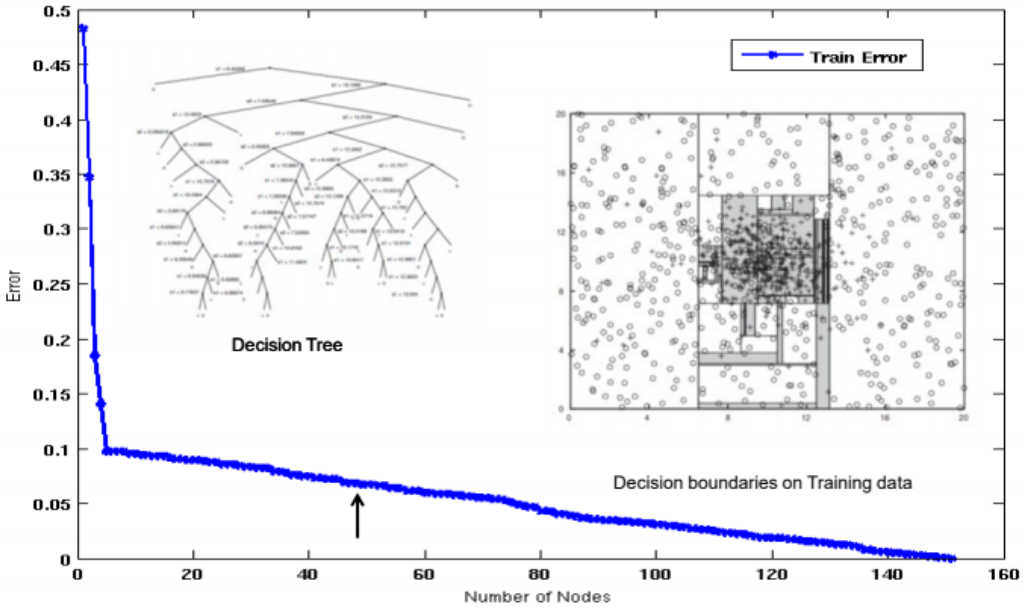
10 % of the data used for training and 90% of the data used for testing

Increasing number of nodes in Decision Trees

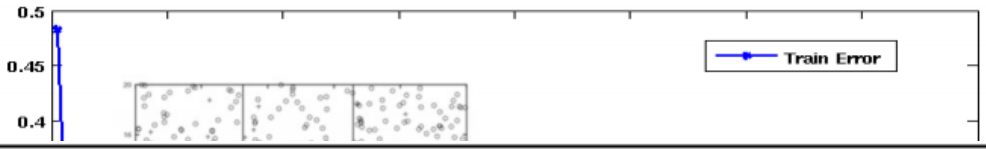
Decision Tree with 4 nodes



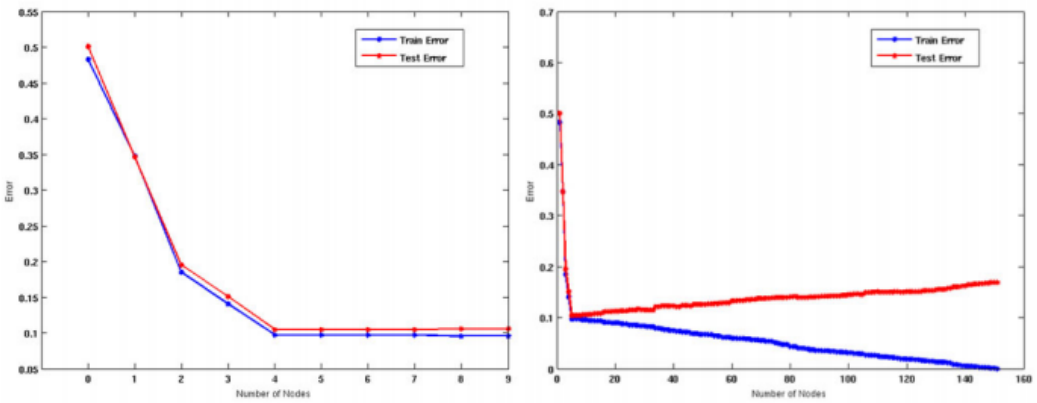
Decision Tree with 50 nodes



Which tree is better?



Model Overfitting

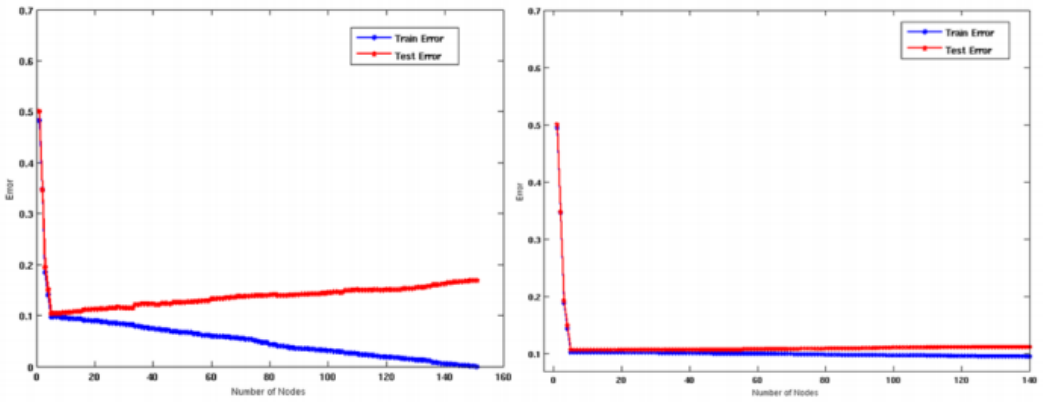


•As the model becomes more and more complex, test errors can start increasing even though training error may be decreasing

Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

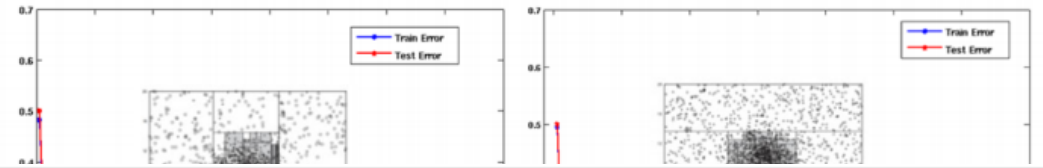
Model Overfitting



Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Model Overfitting



Reasons for Model Overfitting

Limited Training Size

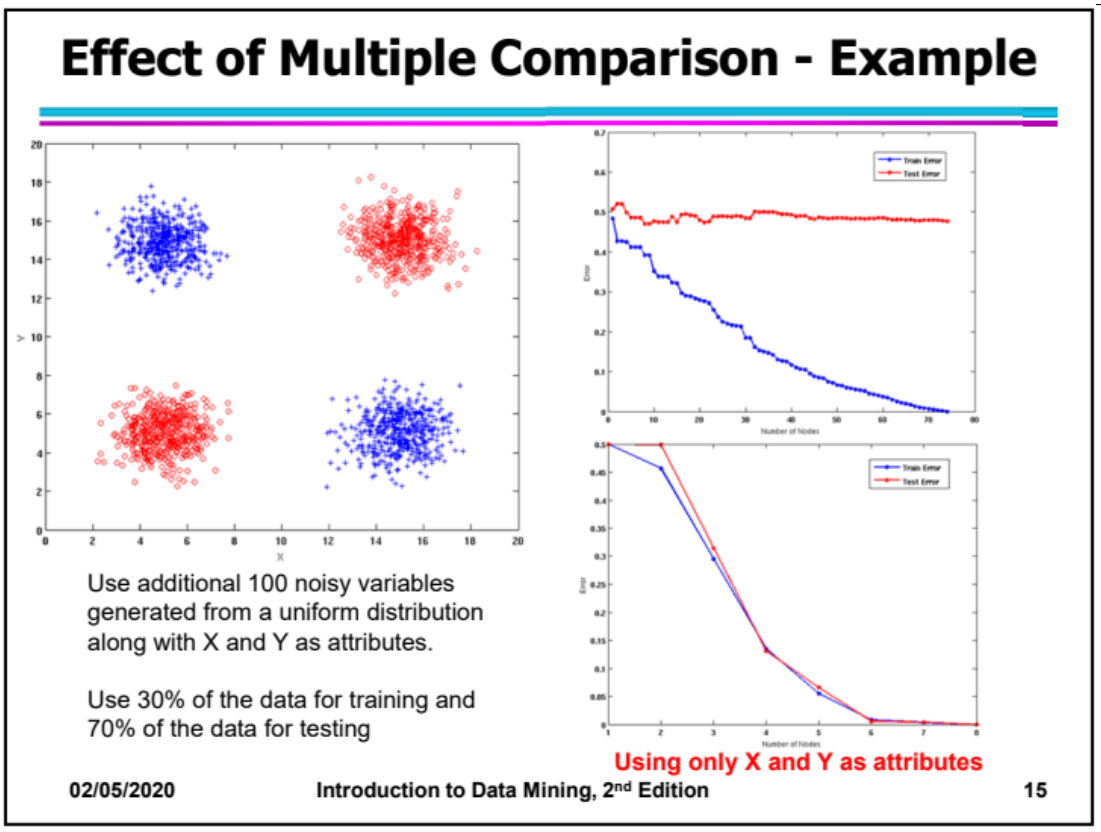
High Model Complexity

- Multiple Comparison Procedure

02/05/2020 数据挖掘导论，第 2 版 12
多重比较程序的效果
考虑预测股市在未来 10 个交易日是否会上涨/下跌的任务
随机猜测:P(正确)= 0.5
连续随机猜 10 次:
第 1 天开始
第 2 天下跌第 3 天下跌第 4 天上涨第 5 天下跌第 6 天下跌第 7 天上涨第 8 天上涨第 9 天上涨第 10 天下跌
0.0547
2
10
10 9
10 8
10
(#) 10
·p·纠正⊠

02/05/2020 数据挖掘导论，第 2 版 13
多重比较程序的效果
方法:
获得 50 名分析师
每位分析师随机猜 10 次，选出最有把握的分析师
正确预测的数量
至少一名分析师做出至少 8 个正确预测的概率
(# 8) 1 (1 0.0547) 0.9399 P 正确⊠⊠⊠

02/05/2020 数据挖掘导论，第 2 版 14
多重比较程序的效果
许多算法采用以下贪婪策略:初始模型:M
替代模型:m ’= m ，
其中⊠是要添加到模型中的组件(例如，决策树的测试条件)保持 M' if 改进，⊠(M,M') >
通常情况下，⊠是从一组可选组件中选出的，⊠ = { 1，⊠2，…，⊠k}
如果有许多替代方案可用，人们可能会无意中向模型中添加不相关的组件，导致模型过度拟合



15

02/05/2020 数据挖掘导论，第 2 版 16
关于过度拟合的注记
过度拟合导致决策树比需要的更复杂
训练误差不能很好地估计树在以前看不到的记录上的表现
需要估算泛化误差的方法

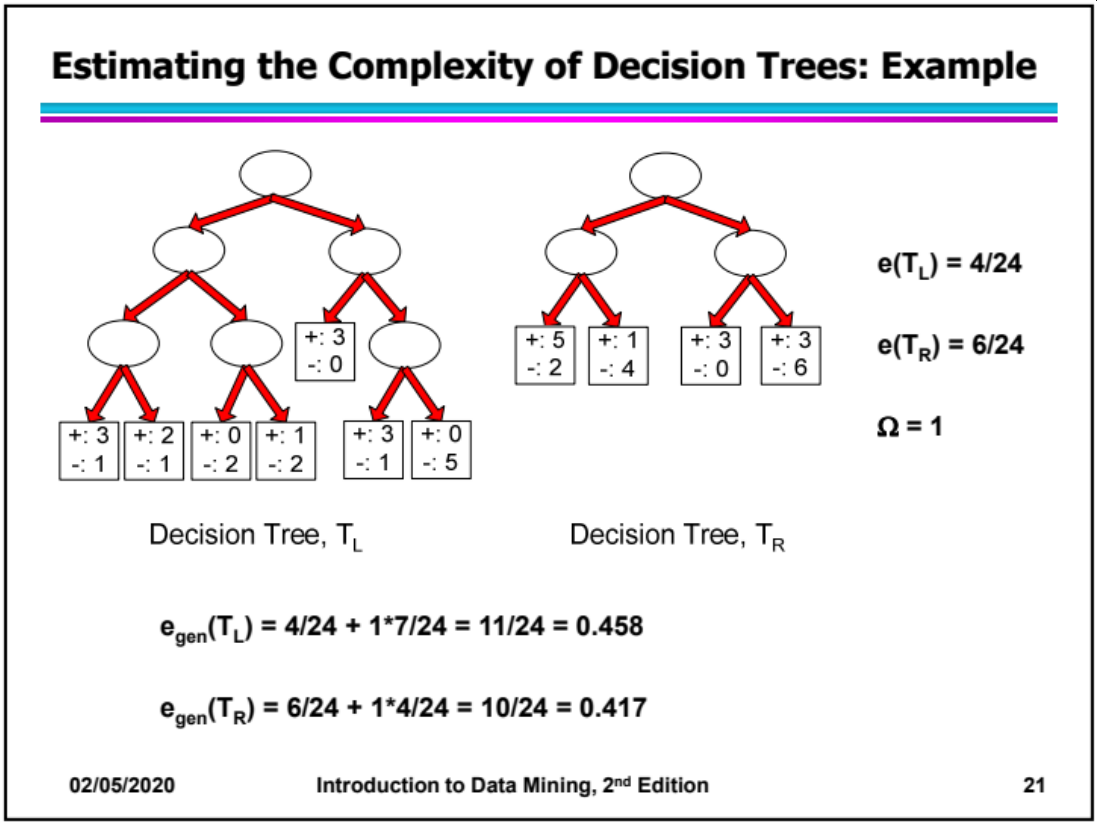
02/05/2020 数据挖掘导论，第 2 版 17
型号选择
在模型构建期间执行目的是确保模型不会过度
复杂(为了避免过度拟合)需要使用验证集来估计泛化误差
结合模型复杂性估计统计界限

02/05/2020 数据挖掘导论，第 2 版 18
型号选择:
使用验证集

将培训数据分为两部分:培训集:
用于模型建造
验证集:
用于估计推广误差⊠注:验证集不同于测试集
缺点:
可用于培训的数据较少

02/05/2020 数据挖掘导论，第 2 版 19
型号选择:
整合模型复杂性
基本原理:奥卡姆剃刀
给定两个具有相似泛化误差的模型，人们应该更喜欢更简单的模型而不是更复杂的模型
一个复杂的模型更有可能被偶然拟合
因此，在评估模型时应该考虑模型的复杂性
一般误差(模型)=列车。错误(型号、列车。数据)+
十复杂性(模型)

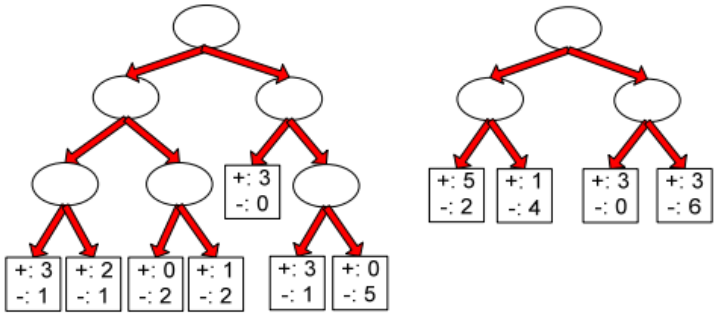
02/05/2020 数据挖掘导论，第二版 20
决策树复杂度的估计
具有 k 个叶节点的决策树的悲观误差估计；
err(T):所有训练记录的错误率⊠:权衡超参数(类似于)
添加叶节点的⊠相对成本
k:叶节点数
培训记录总数



Estimating the Complexity of Decision Trees

Resubstitution Estimate:

- Using training error as an optimistic estimate of generalization error
- Referred to as optimistic error estimate



$e(T_L) = 4/24$

$e(T_R) = 6/24$

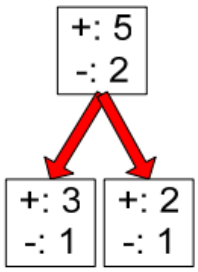
Decision Tree, T_L Decision Tree, T_R

22

02/05/2020 数据挖掘导论，第 2 版 23
最小描述长度
成本(模型，数据)=成本(数据|模型)+ x 成本(模型)成本是编码所需的位数。寻找成本最低的模型。
成本(数据|模型)编码错误分类错误。成本(模型)使用节点编码(子节点数)
加上分裂条件编码。
甲乙
a。
b?
c?
0 1
0
1
是不是
B1 B2
C1·C2
X y
X1 1
X2 0
X3 0
X4 1
… …
Xn 1
X y
X1?
X2?
X3?

X4?
… …
Xn?

Estimating Statistical Bounds



$$e'(N,e,\alpha) = \frac{e + \frac{z_{\alpha/2}^2}{2N} + z_{\alpha/2}\sqrt{\frac{e(1-e)}{N} + \frac{z_{\alpha/2}^2}{4N^2}}}{1 + \frac{z_{\alpha/2}^2}{N}}$$

Before splitting: e = 2/7, e'(7, 2/7, 0.25) = 0.503

e'(T) = 7 × 0.503 = 3.521

After splitting:

$e(T_L)$ = 1/4, e'(4, 1/4, 0.25) = 0.537

$e(T_R)$ = 1/3, e'(3, 1/3, 0.25) = 0.650

e'(T) = 4 × 0.537 + 3 × 0.650 = 4.098

Therefore, do not split

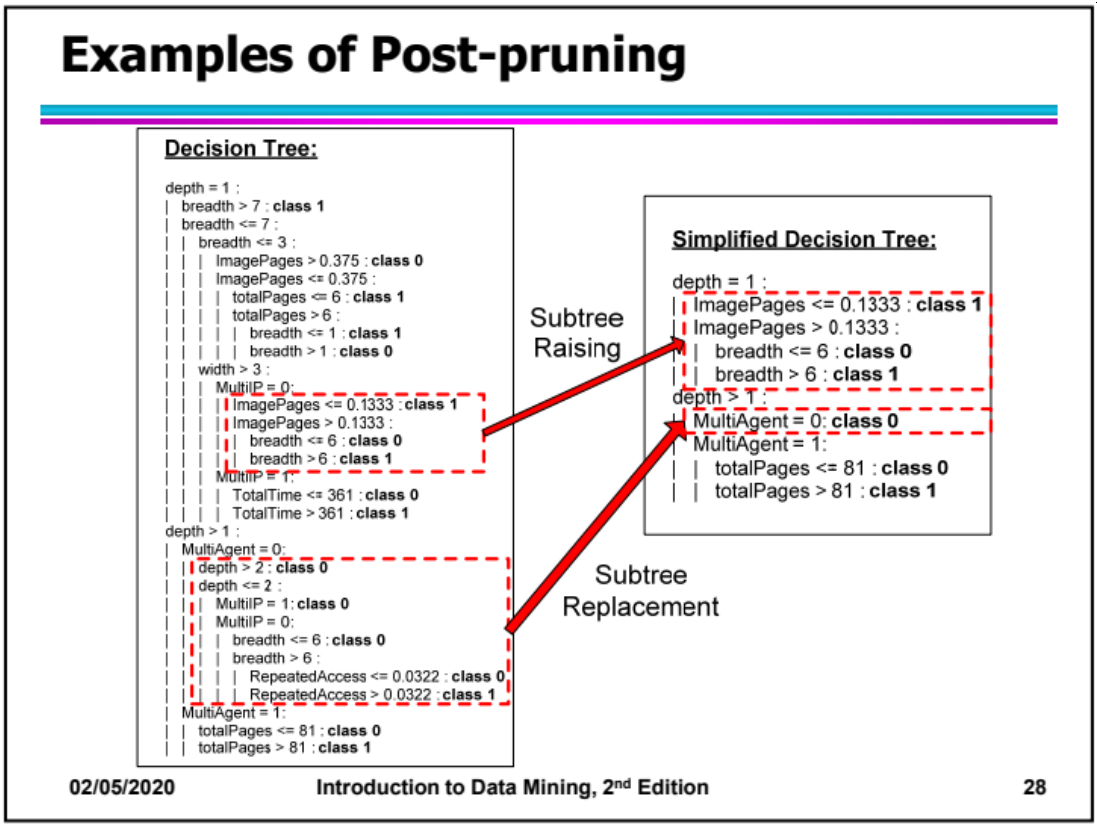
24

02/05/2020 数据挖掘导论，第 2 版 25
决策树的模型选择
预修剪(提前停止规则)
在算法成为完全成熟的树之前停止算法。节点的典型停止条件:
如果所有实例都属于同一个类，□停止；如果所有属性值都相同，□停止
更严格的条件:
如果实例数量少于用户指定的阈值，□停止
如果实例的类别分布独立于可用特征(例如，使用□测试)，则□停止
如果扩展当前节点不能改善杂质度量(例如，基尼系数或信息增益)，则□停止。
如果估计的泛化误差低于某个阈值，□停止

02/05/2020 数据挖掘导论，第 2 版 26
决策树的模型选择
修剪后
将决策树扩展到整个子树替换
以自下而上的方式修剪决策树的节点
如果修剪后泛化误差改善，用叶节点替换子树
叶节点的□类标签由子树中的大多数实例类确定
子树提升

用最常用分支替换子树

02/05/2020 数据挖掘导论，第 2 版 27
后修剪示例
a。
第一等的
A2 A3
A4
等级=是 20 等级=否 10
误差= 10/30
训练误差(分割前)= 10/30 悲观误差= (10 + 0.5)/30 = 10.5/30 训练误差(分割后)= 9/30 悲观误差(分割后)
= (9 + 4 0.5)/30 = 11/30 修剪！
类别=是 8 类别=否 4
类别=是 3 类别=否 4
类别=是 4 类别=否 1
类别=是 5 类别=否 1



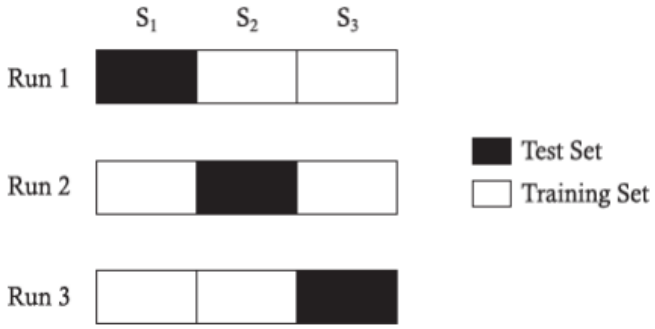
28

02/05/2020 数据挖掘导论，第 2 版 29
模型评估
目的:
评估分类器在以前看不见的数据(测试集)保持上的性能
保留 k%用于训练，保留(100-k)%用于测试随机二次抽样:重复保持交叉验证

将数据划分为 k 个不相交的子集
k-折叠:在 k-1 分区上训练，在剩余的一个分区上测试，留下一个:k=n

Cross-validation Example

3-fold cross-validation



02/05/2020 数据挖掘导论，第 2 版 31
交叉验证的变化
重复交叉验证
多次执行交叉验证给出对
泛化误差分层交叉验证
在培训和测试中保证相同百分比的班级标签
当类不平衡并且样本很小时，这一点很重要
使用嵌套交叉验证方法进行模型选择和评估