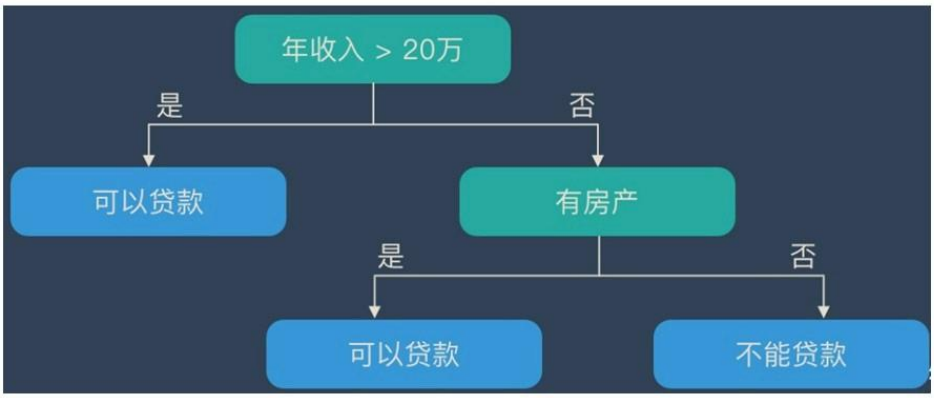
一、集成学习算法概述

中国石油大学 CHINA UNIVERSITY OF PETROLEUM



决策树（Decision tree）

决策树算法

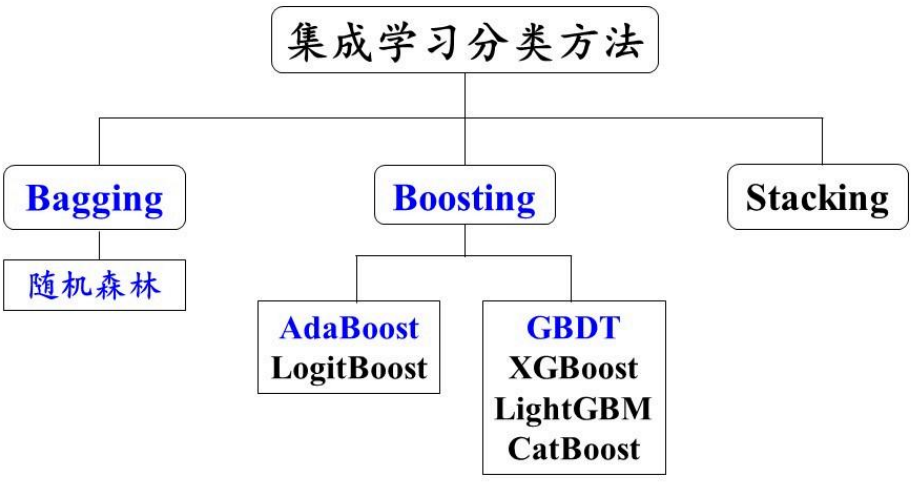
Quinlan 于1986年提出了著名的ID3算法

Breiman等人在1984年提出了CART算法

3

一、集成学习算法概述

中国石油大学 CHINA UNIVERSITY OF PETROLEUM



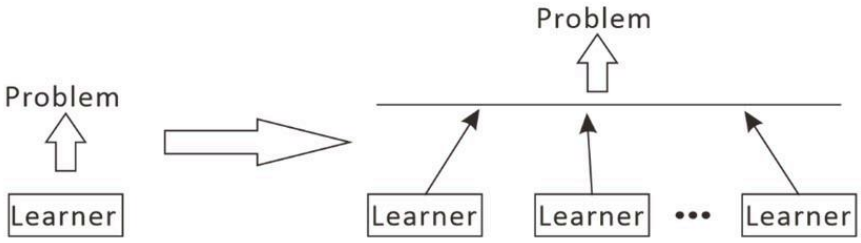
4

一、集成学习算法概述

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

通常使用一个学习器解决一个分类预测问题。

集成学习（Ensemble learning）同时利用多个子学习器来解决问题。



集成学习基本结构图

一、集成学习算法概述

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

1. Bagging（Bootstrap aggregating）算法是于1996年由Breiman等人提出的一种基本并行化集成学习框架。

随机森林算法(Random forest, RF)是由Breiman于2001年在Bagging框架的基础上改进而提出的一种机器学习算法。决策树算法是随机森林算法的子分类器。

Leo Breiman(1928-2005)
加州大学伯克利分校统计学荣誉教授

a man who loved to turn numbers into practical and useful applications

一、集成学习算法概述

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

2. Boosting算法是由Schapire于1990年提出的一种串行化集成学习框架。常用有AdaBoost和GBMs两大类算法。

Robert Schapire

普林斯顿大学
计算机科学系前教授

现就职于微软

- **(1) AdaBoost 是于1997年由Freund and Schapire 提出的一种通过增加错误分类数据点的权重来改变模型的缺陷的Boosting算法。**

Yoav Freund

加州大学圣地亚哥分校
计算机科学教授

一、集成学习算法概述　　2. Boosting算法

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

- **(2) GBMs算法是一种通过在训练中对损失函数的残差来进行梯度优化的Boosting算法框架，包括GBDT、XGBoost等算法。**

✓ **GBDT算法是由Friedman于1999年提出的基于决策树的算法。**

Jerome Friedman

斯坦福
大学
统计系
教授

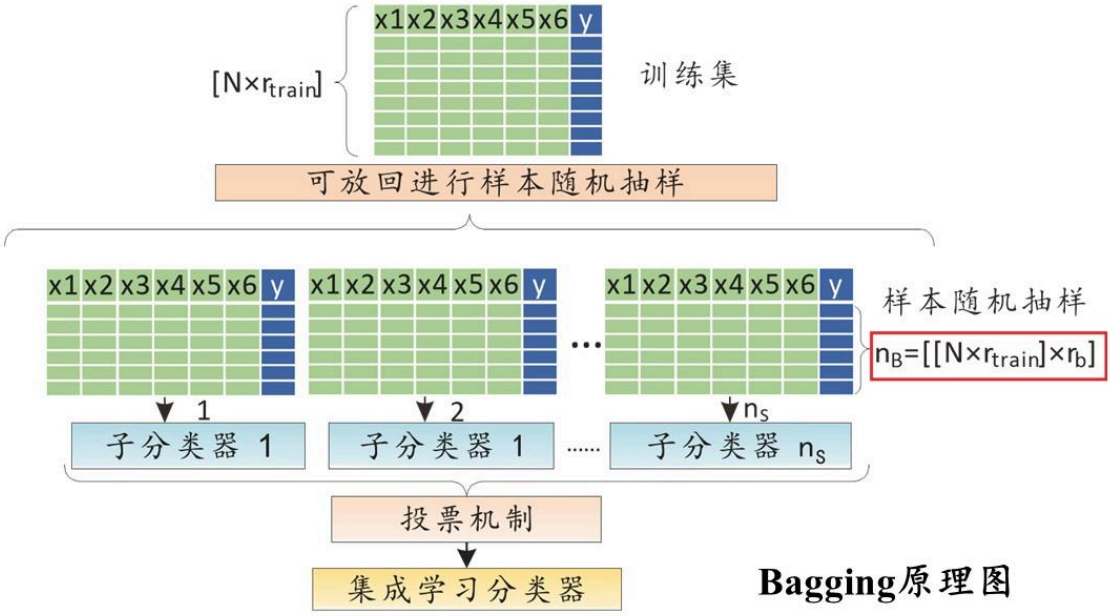
✓ **XGBoost算法是Chen于2016年提出的基于GBDT的改进算法。**

陈天奇
卡内基梅隆大学
CMU助理教授

华盛顿大学计算
机系博士
上海交大本科

二、Bagging集成学习算法　1. Bagging集成算法

该方法是一种并行化训练子分类器的集成学习方法。

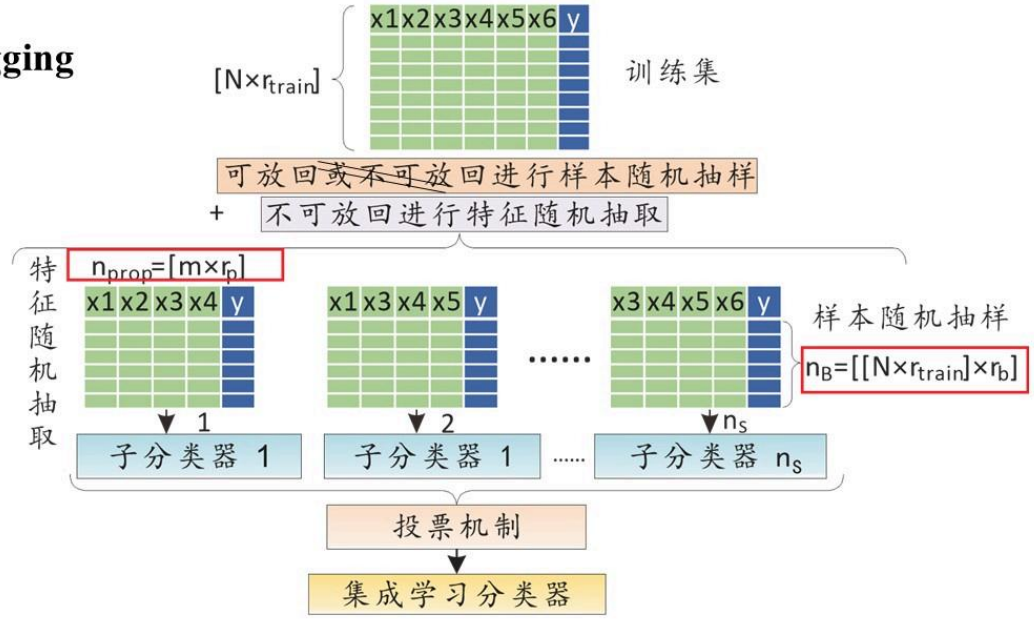


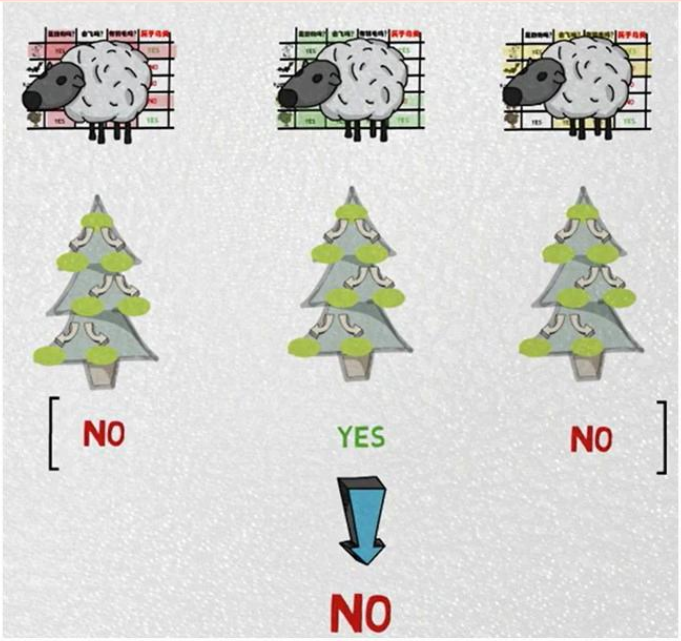
Bagging原理图

二、Bagging集成学习算法　2.随机森林

该方法为Bagging的扩展方法

- 不仅包括样本随机抽样
- 还增加了特征随机抽取

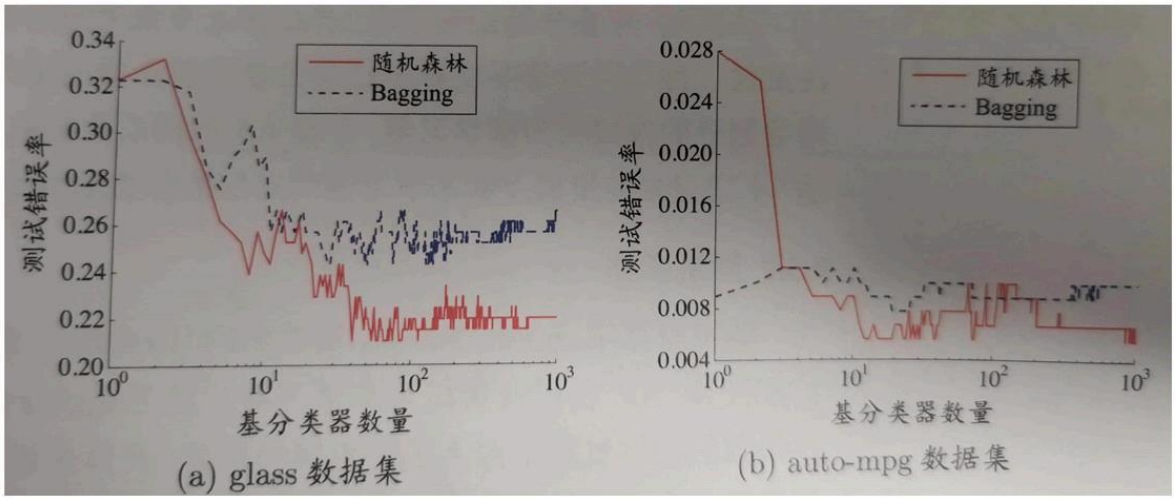




二、Bagging集成学习算法　2.随机森林

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

随机森林相比Bagging只做了小的改动，
随机森林除了继承了Bagging的样本扰动，还来自于属性扰动



12

CART决策树应用实例

例，考察CNL值为28.8进行划分，根据CART划分规则进行计算。

当CNL值>28.8时，分类标签有两种。

$$GINI(CNL>28.8)=1-(\frac{1}{2})^2-(\frac{1}{2})^2=\frac{1}{2}$$

当CNL值≤28.8时，分类标签有一种。

$$GINI(CNL\leq 28.8)=1-(1)^2=0$$

GINI(根据CNL按照28.8进行划分)

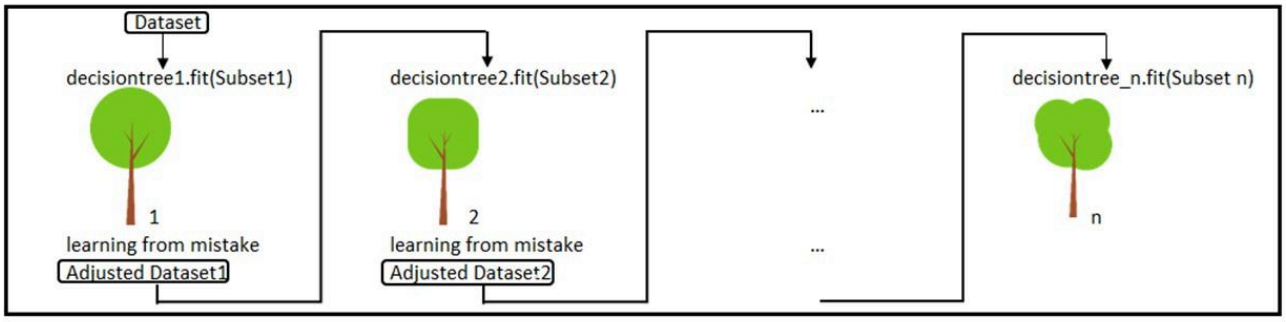
$$=\frac{2}{3}\times\frac{1}{2}+\frac{1}{3}\times 0=\frac{1}{3}$$

No.	DEN	CNL	岩性
1	1.79	40.6	C
2	1.98	37.6	C
3	2.21	36.1	C
4	2.24	39.0	M
5	2.27	42.1	M
6	2.34	43.1	M
7	2.53	21.5	PS
8	2.52	20.8	PS
9	2.51	20.1	PS

17

Boosting基于通过下一个学习器来改善上一个学习器的错误的串行化原则来构建集成学习分类器

Boosting框架基本可以分为AdaBoost 和 GBMs



Boosting

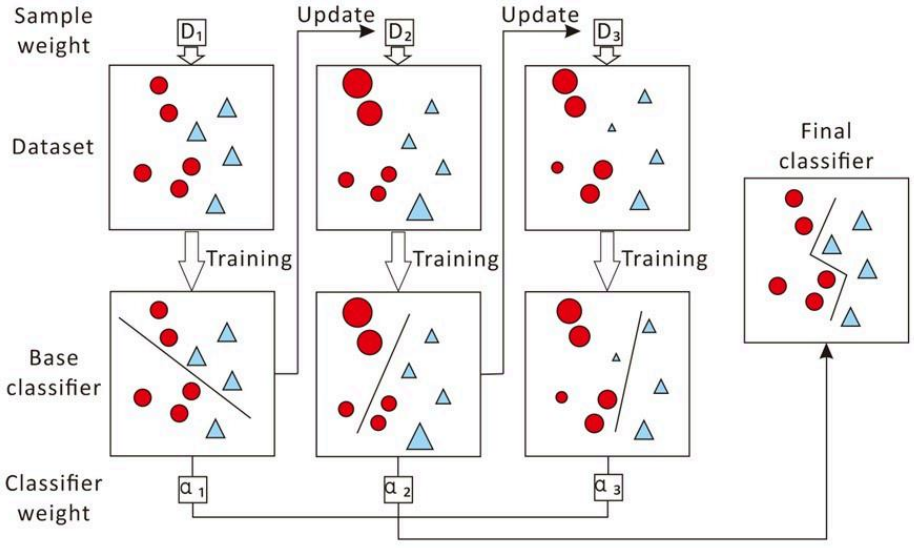
25

三、Boosting集成学习算法

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

1. AdaBoost

根据基学习器的表现对训练样本分布进行调整，使得之前基学习器做错的训练样本在后续受到更多关注，然后基于调整后的样本分布来训练下一个基学习器。



AdaBoost原理图

三、Boosting集成学习算法

中国石油大学 CHINA UNIVERSITY OF PETROLEUM

1.AdaBoost优缺点及总结

优点： 算法简单，精度高泛化能力强，不容易过拟合。

缺点： 受异常点影响大，时间成本高。

算法根据新的分布$\mathcal{D}_t$来最小化分类误差，并要求误差小于**0.5**，该思想本质为对每个样本的分布加权重以实现残差逼近。