数据库系统概论新技术篇

数据挖掘

李翠平

中国人民大学信息学院

概览

- 什么是分类？
- 分类 vs. 预测
- 分类的过程
  - 建立模型阶段
  - 使用模型阶段
- 分类过程用到的三类数据集
- 常用的分类方法
- 一种典型的分类方法：决策树分类

什么是分类

- 属于模型挖掘，更确切地讲，是预测建模的过程
- 预测建模的目的是，根据观察到的对象特征值预测它的其他特征值
- 例如，根据年龄、收入、是否学生、信用情况这四个方面的特征，预测某个人是否会购买笔记本电脑
- 这是一个典型的分类问题

分类 vs. 预测

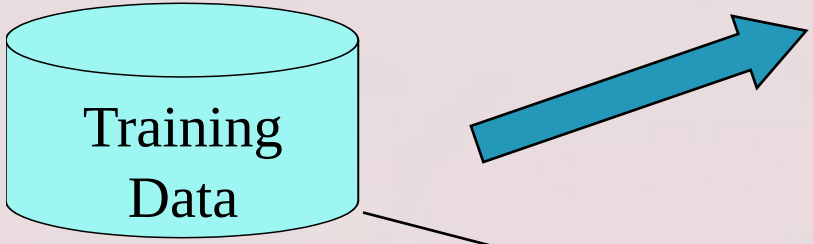
- 分类
  - 构造、使用模型来对某个样本的类别进行判别
  - 主要用于对离散的数据进行预测
  - 典型应用：信誉评估、医学诊断

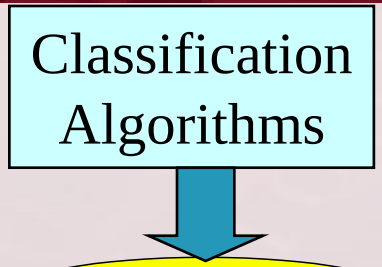
- 预测
  - 构造、使用模型来对某个样本的值进行估计，例如预测某个不知道的值或者缺失值
  - 主要用于对连续或有序的数据进行预测
  - 典型应用：性能预测

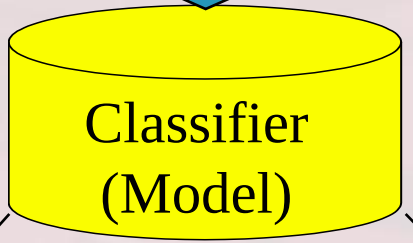
分类的过程

- ❖ 第一步，建立模型阶段
  - ➢ 用来构造模型的数据集被称为训练集
  - ➢ 模型一般表示为：分类规则，决策树或者数学公式
- ❖ 第二步，使用模型阶段
  - ➢ 首先测试模型的准确性
    - 用测试集和由模型进行分类的结果进行比较
    - 两个结果相同所占的比率称为准确率
    - 测试集和训练集必须不相关
  - ➢ 如果准确性可以接受的话，使用模型对新数据进行分类

分类过程：建立模型阶段



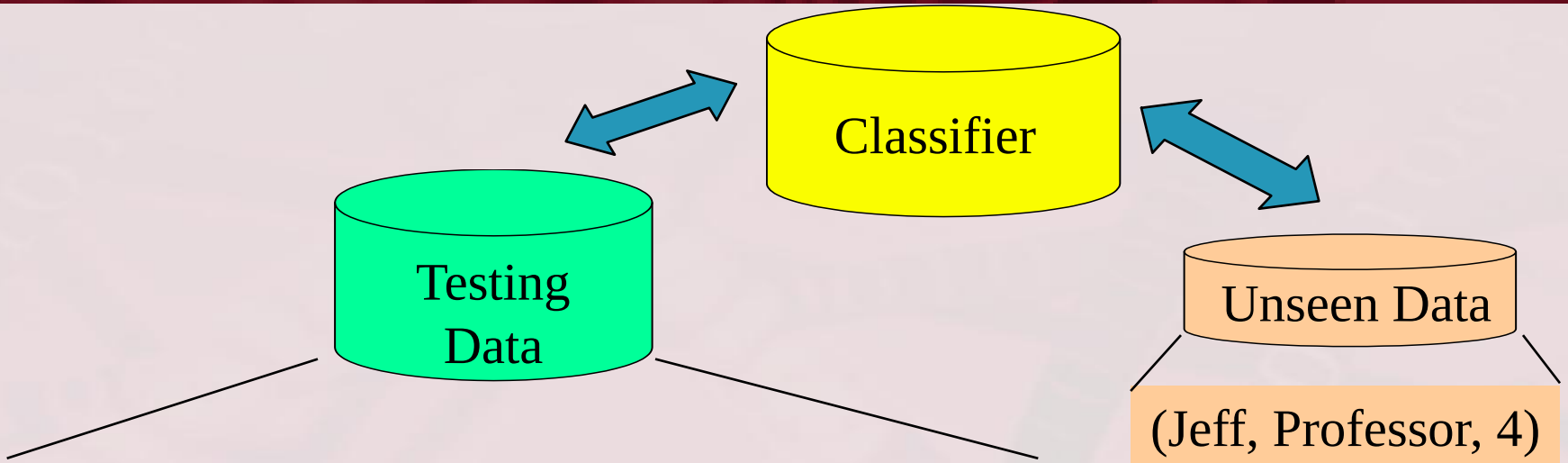




NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

分类过程：使用模型阶段



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

分类过程用到的三类数据集

训练集

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

测试集

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

新数据

NAME	RANK	YEARS	TENURED
Jeff	Professor	4	?

分类方法进行评价

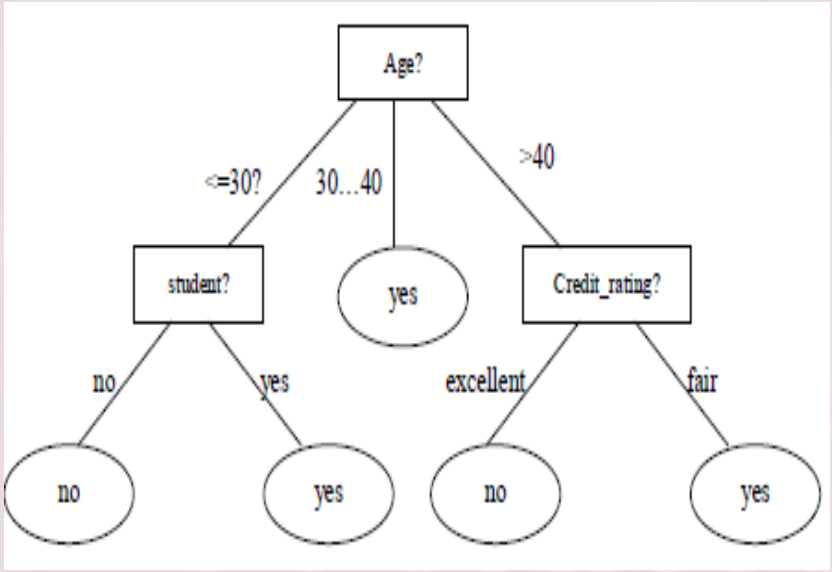
- 准确性
- 速度
  - 构造模型的时间 (训练时间)
  - 使用模型的时间 (预测时间)
- 鲁棒性
  - 能够处理噪声和缺失数据
- 可伸缩性
  - 对磁盘级的数据库有效
- 易交互性
  - 模型容易理解，具有较好的洞察力

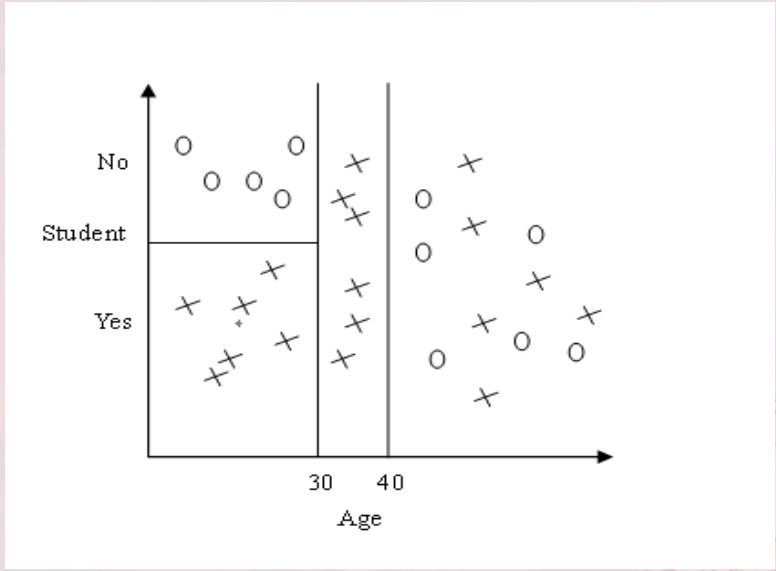
常用的分类方法

- 决策树分类
- 贝叶斯分类
- 支持向量机
- 神经网络
- K近邻分类

决策树分类

决策树分类的主要任务是要确定各个类别的决策区域，或者说，确定不同类别之间的边界。在决策树分类模型中，不同类别之间的边界通过一个树状结构来表示





决策树算法的宏观思考

- 最大高度 = 决策属性的个数
- 树越矮越好
- 要把重要的好的属性放在树根

因此，决策树建树算法就是： 选择树根的过程

决策树分类

1 开始时，所有的训练集样本都在树根

2 属性都是可分类的属性(如果是连续值的话，先要对其进行离散化)

停止划分的条件：

1 某个节点上的所有样本都属于相同的类别

2 所有属性都用到了– 采用多数有效法对叶子节点分类

3 没有样本了

决策树分类第一步：选择属性，作为树根

- 比较流行的属性选择方法：信息增益
- 信息增益最大的属性被认为是最好的树根

属性选择方法：信息增益计算

- 用S表示训练集，假设分类属性具有m个不同的值，也就是说共有m个不同的分类 $C_i(i = 1, \ldots, m)$, 用s_i 表示S中属于分类Ci的样本的个数

- 则信息收益可以用如下三步求出
  - 求information:

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

  - 对每个属性求entropy, 假设属性A的值为$\{a_1,a_2,\ldots,a_v\}$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

  - 对每个属性求information gain:

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A)$$

属性选择方法：信息增益计算示例

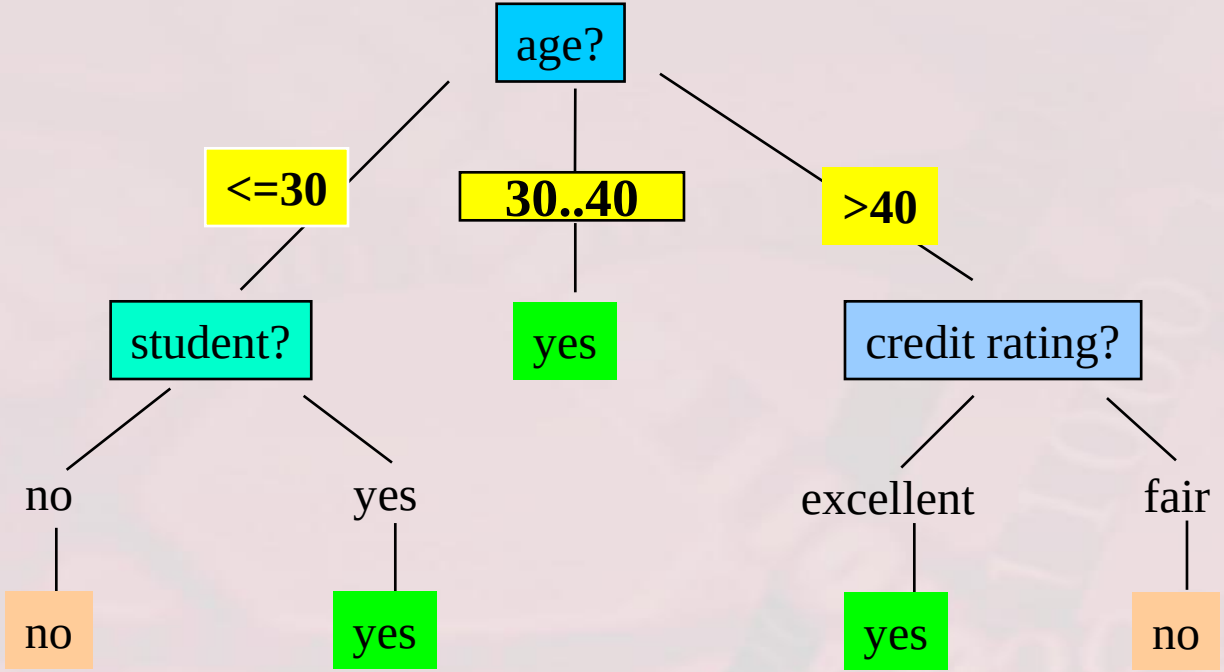
共有5个属性

前4个属性用作预测属性，最后一个属性是类别属性

共有14个样本，或者说14条记录

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

属性选择方法：信息增益计算示例



$$I(p,n) = I(\frac{9}{14},\frac{5}{14}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

- Class N: buys_computer = "no"
- I(p, n) = I(9, 5) =0.940
- Compute the entropy for *age*:

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
30…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$E(age) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = I(p,n) - E(age) = 0.246$$

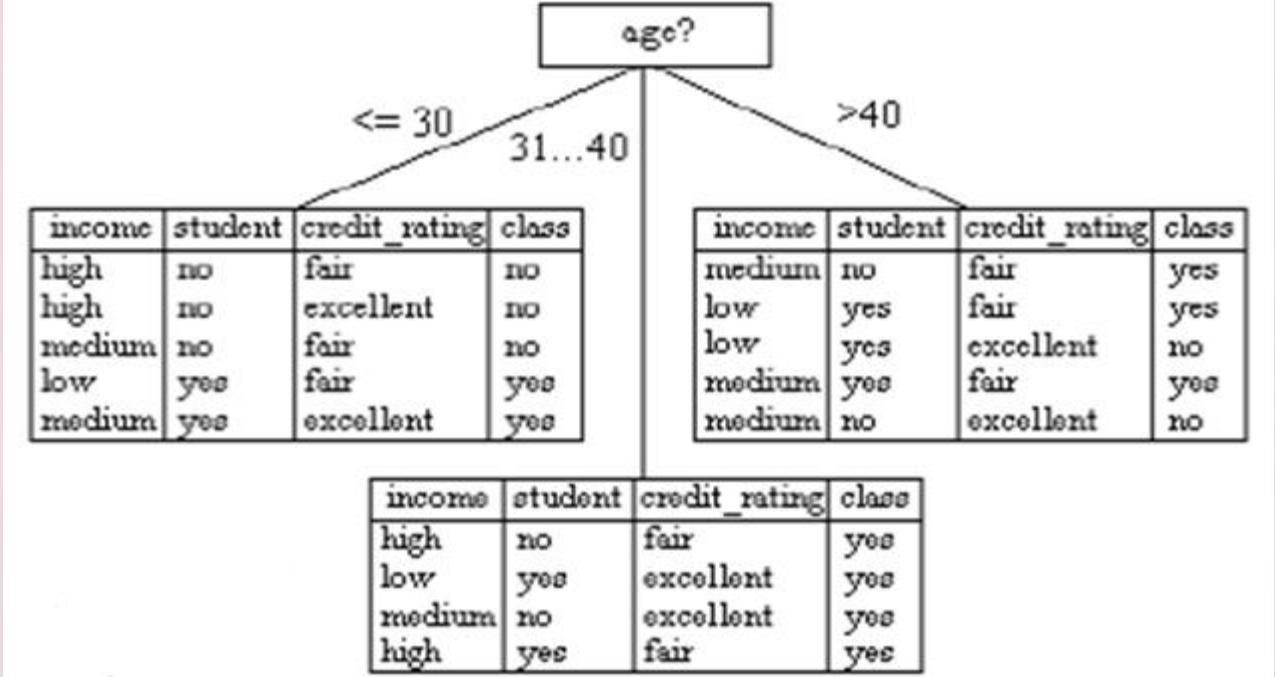
Similarly,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

根据树根，划分训练集



income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

根据属性 age 进行数据集划分

从决策树中抽取决策规则

- 决策树中所蕴含的知识可以表达成IF-THEN规则的形式
- 从根到叶的一条路径生成一条规则
- 路径上的属性值由AND连接起来，构成IF部分
- 叶子节点构成THEN部分，指出所属的分类
- **Example**

IF *age* = “<=30” AND *student* = “*no*” THEN *buys_computer* = “*no*”

IF *age* = “<=30” AND *student* = “*yes*” THEN *buys_computer* = “*yes*”

IF *age* = “31…40” THEN *buys_computer* = “*yes*”

IF *age* = “>40” AND *credit_rating* = “*excellent*” THEN *buys_computer* = “*yes*”

IF *age* = “<=30” AND *credit_rating* = “*fair*” THEN *buys_computer* = “*no*”

谢 谢！