**1.知识发现：**

A、数据清理（消除噪声和删除不一致数据）。

B、数据集成（多种数据源可以组合在一起）。

C、数据选择（从数据库中提取与分析任务相关的数据）。

D、数据变换（通过汇总或聚集操作，把数据变换和统一成适合挖掘的形式）。

E、数据挖掘（基本步骤，使用智能方法提取数据模式）。

F、模式评估（根据某种兴趣度度量，识别代表知识的真正有趣的模式。

G、知识表示（使用可视化和知识表示技术，向用户提供挖掘的知识）。

**2.数据预处理：**

数据清洗：填充缺失值, 识别/去除离群点, 光滑噪音, 并纠正数据中的不一致

数据集成：多个数据库, 数据立方体, 或文件的集成（合并多个数据源中的数据，存入一个一致的数据存储中）

数据规约：得到数据的归约表示, 它小得多, 但产生相同或类似的分析结果：维度规约、数值规约、数据压缩

数据变换：规范化和聚集

数据预处理解决的问题

数据清洗：去除脏数据（可以具体点，比如什么样的脏数据，包括离群数据）

数据集成：合并多个数据源中的数据，存入一个一致的数据存储中。数据规约：数据归约（消减）技术用于帮助从原有庞大数据集中获得一个精简的数据集合，并使这一精简数据集保持原有数据集的完整性。这样在精简数据集上进行数据挖掘显然效率更高，并且挖掘出来的结果与使用原有数据集所获得结果基本相同。

数据集成涉及的问题

主要涉及模式集成（在中介模式与源数据模式上建立映射关系）、属性冗余（可以分局协相关系数、视觉散点图、卡方检验等方法判断两个属性之间的相关性）、实体识别、去除重复数据、数据值冲突的检测与处理（表示、比例或编码不同）

数据归约的常见方法

数据立方体聚集、维规约、数值规约

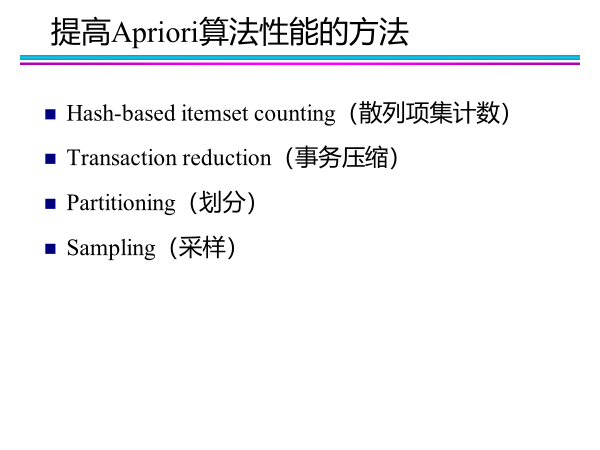
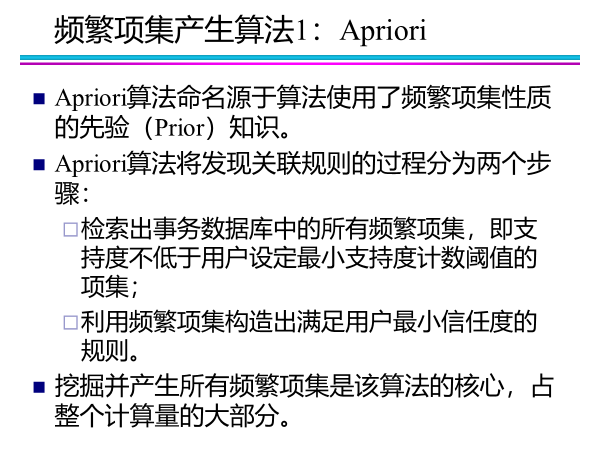
参数方法：回归和对数-线性模型就是例子

非参数方法包括直方图、聚类、抽样和数据立方体聚集

数据压缩：使用变换，以便得到原数据的归约或“压缩”表示，分为有损和无损压缩，维归约和数值归约也可以视为某种形式的数据压缩。

数据转换常用的技术  
光滑（去除噪声，如分箱、聚类、回归）、聚集（avg、sum）、属性构造（长和宽构造面积）、数据概化（泛化）、规范化（最大最小规范化、Z-score规范化（又叫零-均值规范化，数据与均值的差再比上数据集的方差）、小数定标规范化（移动小数点）

**3.关联分析**



优点：

1) Aprioi算法采用逐层搜索的迭代方法，算法简单明了，没有复杂的理论推导，也易于实现。

2) 数据采用水平组织方式

3)采用Apriori 优化方法

4)适合事务数据库的关联规则挖掘。

5)适合稀疏数据集:根据以往的研究,该算法只能适合稀疏数据集的关联规则挖掘，也就是频繁项目集的长度稍小的数据集。

缺陷：

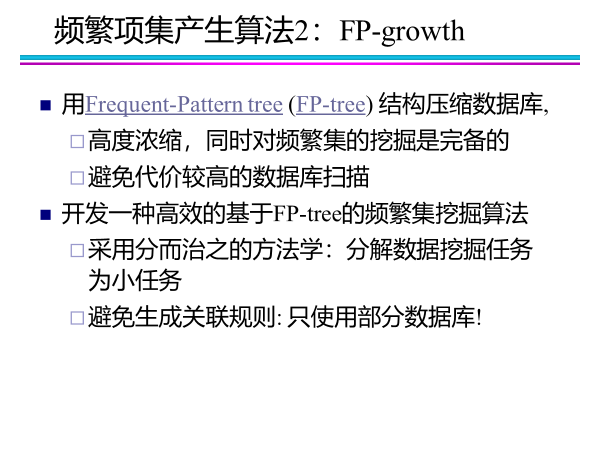
1)对数据库的扫描次数过多。

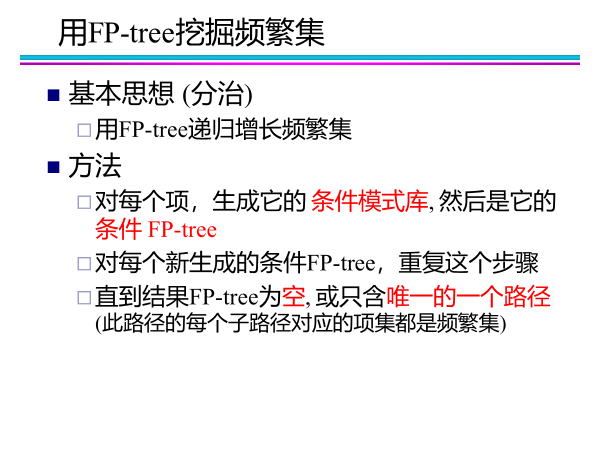
2) Apion算法可能产生大量的候选项集。

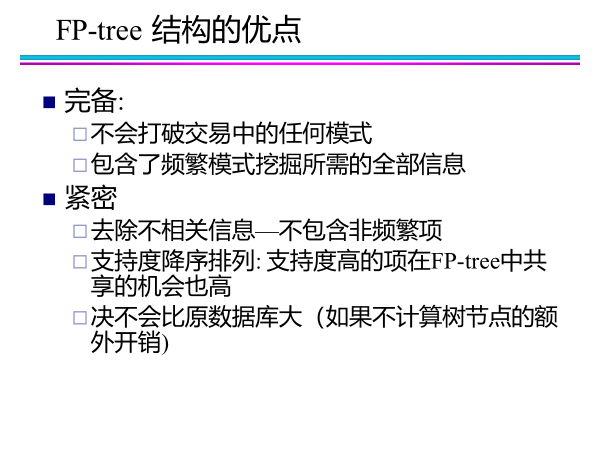
3)在频繁项目集长度变大的情况下,运算时间显著增加。

4)采用唯一支持度,没有考虑各个属性重要程度的不同。

5)算法的适应面窄。







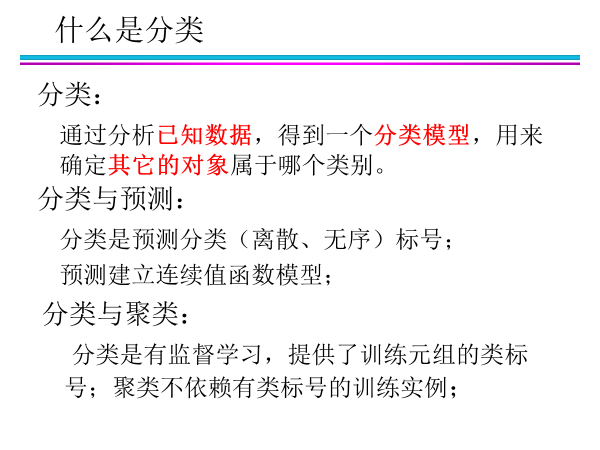
优点：

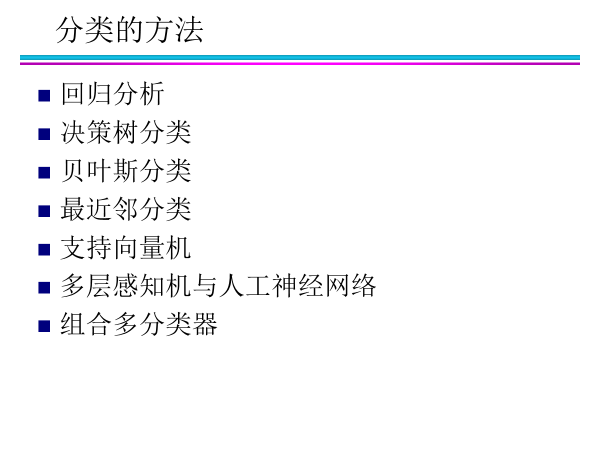
FP-tree是一个高度压缩的结构，它存储了用于挖掘频繁项集的全部信息。

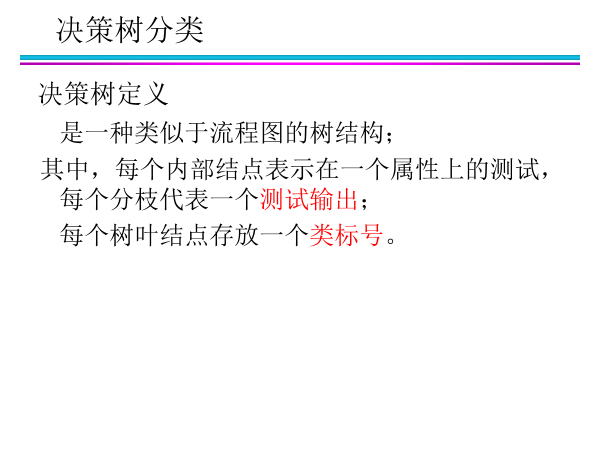
缺点：

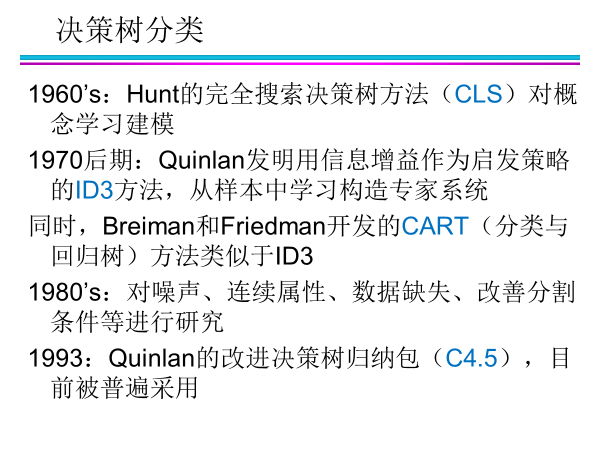
树的子节点过多，例如生成了只包含前缀的树，那么也会导致算法效率大幅度下降。FP-Growth算法需要递归生成条件数据库和条件FP-tree,所以内存开销大，而且只能用于挖掘单维的布尔关联规则。

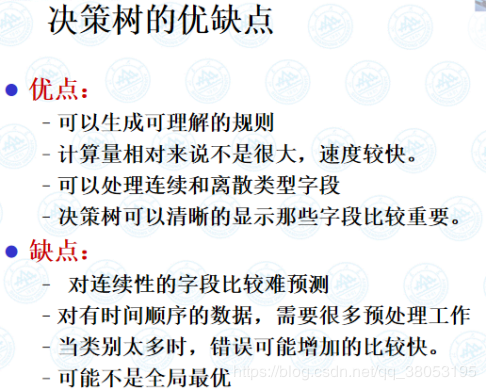
**4.分类**

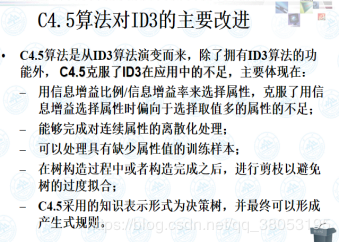
****

****

****

****





ID3

D3算法十分简单，核心是根据“最大信息熵增益”原则选择划分当前数据集的最好特征，信息熵是信息论里面的概念，是信息的度量方式，不确定度越大或者说越混乱，熵就越大。在建立决策树的过程中，根据特征属性划分数据，使得原本“混乱”的数据的熵(混乱度)减少，按照不同特征划分数据熵减少的程度会不一样。在ID3中选择熵减少程度最大的特征来划分数据（贪心），也就是“最大信息熵增益”原则。

同时这是最早提出的一种决策树方法，使用上述信息增益的方式建立。

缺点：只能处理离散型属性，并且对倾向于选择取值较多的属性

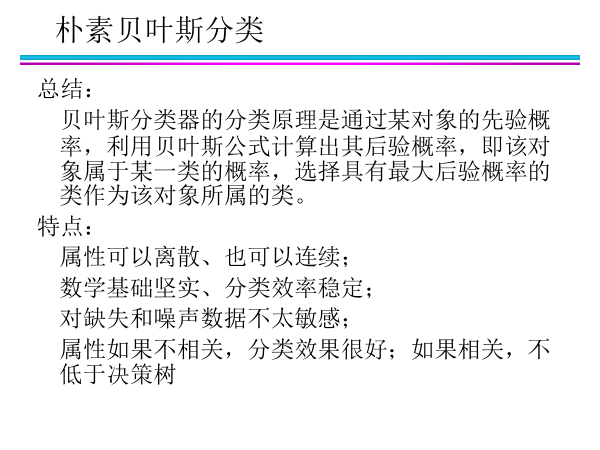
原因：信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大。

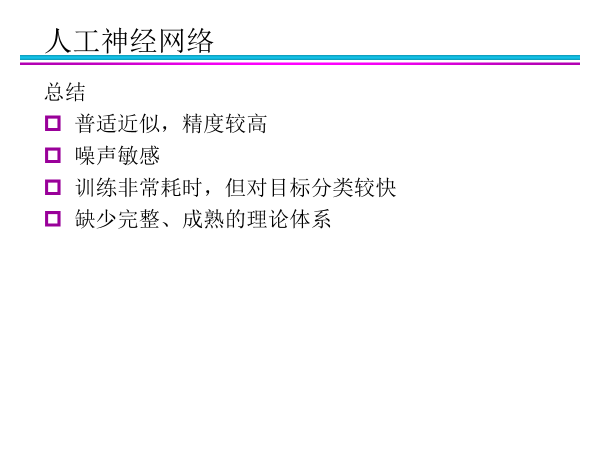
C4.5

C4.5算法流程与ID3相类似，只不过将信息增益改为信息增益比，以解决偏向取值较多的属性的问题，另外它可以处理连续型属性。

K折交叉验证，将初始采样（样本集X，Y）分割成K份，一份被保留作为验证模型的数据（test set），其他K-1份用来训练（train set）。交叉验证重复K次，每份验证一次，平均K次的结果或者使用其它结合方式，最终得到一个单一估测。

这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10折交叉验证是最常用的。（切记每次作为验证模型的数据是不同的）。





**5.聚类**

