机器学习与数据挖掘

一、概念

1. 什么是机器学习？

机器学习是研究如何使用机器来模拟人类活动的一门学科。稍微严格的提法是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。

2. 什么是数据挖掘？

数据挖掘是从大量的数据中挖掘出隐含的、未知的、用户可能感兴趣的和对决策有潜在价值的知识和规则。简单的说，数据挖掘就是从大量的数据中发现有用的信息。

3. 数据挖掘的过程 ？

理解数据和数据的来源、获取相关知识与技术、整合与检查数据、去除错误或不一致的数据、建立模型和假设、实际数据挖掘工作、测试和验证挖掘结果、解释和应用。

4. 与数据挖掘相关的期刊或会议 ？

期刊：ACM TKDD、DMKD、IEEE TKDE

会议: SigKDD、ICDM、SDM

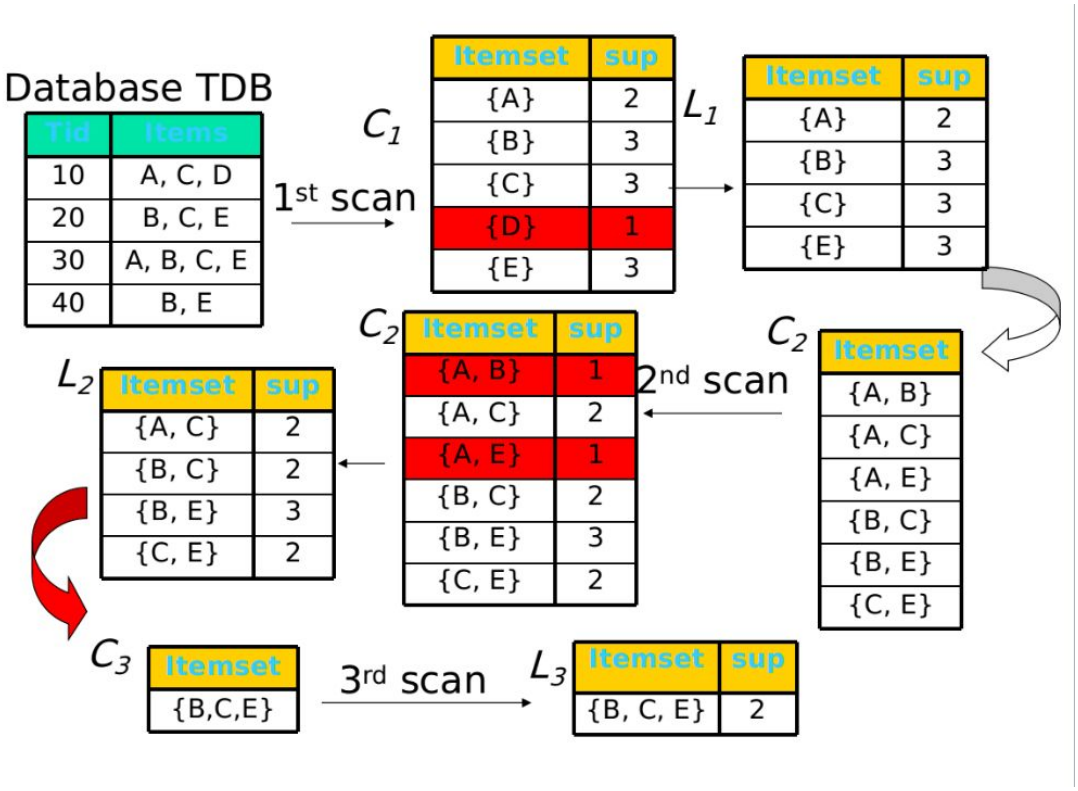
工具: Weka、Rapid Miner(Yale)、illiMine

二、关联规则

a) Apriori 频繁集

注意: sup 表示支持度；

L2 --> C3 不存在{A, C, E}、{A, B, C}是因为在 C2 频繁集中，{A, E}、{A, B}不满足最小支持度。



b) fp-tree 算法

具体过程:

1. 扫描数据库一次，得到频繁 1-项集
2. 把项按支持度递减排序
3. 再一次扫描数据库，建立 FP-tree

TID	Items Bought	(Ordered) Frequent Items
100	a, b, c, d, e, f, g, h	a, b, d, e, f, g
200	a, f, g	a, f, g
300	b, d, e, f, j	b, d, e, f
400	a, b, d, i, k	a, b, d
500	a, b, e, g	a, b, e, g

Threshold = 3

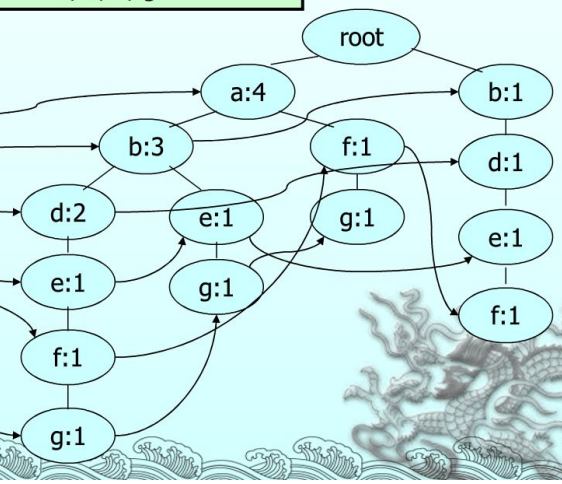
Item	Frequency
a	4
b	4
c	1
d	3
e	3
f	3
g	3
h	1
i	1
j	1
k	1

Item	Frequency
a	4
b	4
d	3
e	3
f	3
g	3

TID	Items Bought	(Ordered) Frequent Items
100	a, b, c, d, e, f, g, h	a, b, d, e, f, g
200	a, f, g	a, f, g
300	b, d, e, f, j	b, d, e, f
400	a, b, d, i, k	a, b, d
500	a, b, e, g	a, b, e, g

Threshold = 3

Item	Head of node-link
a	
b	
d	
e	
f	
g	



三、分类

四、聚类

K-means（k-均值）步骤：

1. 从数据中随机抽取 K 个点作为初始聚类的中心，由这个中心代表各个聚类；

2. 计算数据中心中所有的点到这 K 个点的距离，将点归到离其最近的聚类里；

3. 调整聚类的中心，即将聚类的中心移动到聚类的几何中心(即平均值)处，也就是 K-means 中的 mean 的含义；

4. 重复第 2、3 步知道聚类的中心不再移动，此时算法收敛。

K-means（k-中心点）步骤：