数据挖掘——概念概念与技术
Data Mining
Concepts and Techniques

习题解答

Jiawei Han　Micheline Kamber　著
范明　孟晓峰 译

目录

第 1 章 引言

1.1 什么是数据挖掘？在你的回答中，针对以下问题：

1.2 1.6 定义下列数据挖掘功能：特征化、区分、关联和相关分析、预测聚类和演变分析。使用你熟悉的现实生活的数据库，给出每种数据挖掘功能的例子。

解答：

- **特征化**是一个目标类数据的一般特性或特性的汇总。例如，学生的特征可被提出，形成所有大学的计算机科学专业一年级学生的轮廓，这些特征包括作为一种高的年级平均成绩(GPA：Grade point aversge)的信息，还有所修的课程的最大数量。
- **区分**是将目标类数据对象的一般特性与一个或多个对比类对象的一般特性进行比较。例如，具有高 GPA 的学生的一般特性可被用来与具有低 GPA 的一般特性比较。最终的描述可能是学生的一个一般可比较的轮廓，就像具有高 GPA 的学生的 75%是四年级计算机科学专业的学生，而具有低 GPA 的学生的 65%不是。
- **关联**是指发现关联规则，这些规则表示一起频繁发生在给定数据集的特征值的条件。例如，一个数据挖掘系统可能发现的关联规则为：major(X, “computing science”) ⇒ owns(X, “personal computer”) [support=12%, confidence=98%]
  其中，X 是一个表示学生的变量。这个规则指出正在学习的学生，12%（**支持度**）主修计算机科学并且拥有一台个人计算机。这个组一个学生拥有一台个人电脑的概率是 98%（置信度，或确定度）。
- **分类**与**预测**不同，因为前者的作用是构造一系列能描述和区分数据类型或概念的模型（或功能），而后者是建立一个模型去预测缺失的或无效的、并且通常是数字的数据值。它们的相似性是他们都是预测的工具：分类被用作预测目标数据的类的标签，而预测典型的应用是预测缺失的数字型数据的值。

- **聚类**分析的数据对象不考虑已知的类标号。对象根据最大花蕾内部的相似性、最小化类之间的相似性的原则进行聚类或分组。形成的每一簇可以被看作一个对象类。聚类也便于分类法组织形式，将观测组织成类分层结构，把类似的事件组织在一起。
- **数据延边分析**描述和模型化随时间变化的对象的规律或趋势，尽管这可能包括时间相关数据的特征化、区分、关联和相关分析、分类、或预测，这种分析的明确特征包括时间序列数据分析、序列或周期模式匹配、和基于相似性的数据分析

1.3 1.9 列举并描述说明数据挖掘任务的五种原语。

解答：

用于指定数据挖掘任务的五种原语是：

- **任务相关数据：**这种原语指明给定挖掘所处理的数据。它包括指明数据库、数据库表、或数据仓库，其中包括包含关系数据、选择关系数据的条件、用于探索的关系数据的属性或维、关于修复的数据排序和分组。
- **挖掘的数据类型：**这种原语指明了所要执行的特定数据挖掘功能，如特征化、区分、关联、分类、聚类、或演化分析。同样，用户的要求可能更特殊，并可能提供所发现的模式必须匹配的模版。这些模版或超模式（也被称为超规则）能被用来指导发现过程。
- **背景知识：**这种原语允许用户指定已有的关于挖掘领域的知识。这样的知识能被用来指导知识发现过程，并且评估发现的模式。关于数据中关系的概念分层和用户信念是背景知识的形式。
- **模式兴趣度度量：**这种原语允许用户指定功能，用于从知识中分割不感兴趣的模式，并且被用来指导挖掘过程，也可评估发现的模式。这样就允许用户限制在挖掘过程返回的不感兴趣的模式的数量，因为一种数据挖掘系统可能产生大量的模式。兴趣度测量能被指定为简易性、确定性、适用性、和新颖性的特征。
- **发现模式的可视化：**这种原语述及发现的模式应该被显示出来。为了使数据挖掘能有效地将知识传给用户，数据挖掘系统应该能将发现的各种形式的模式展示出来，正如规则、表格、饼或条形图、决策树、立方体

或其它视觉的表示。

1.4 1.13 描述以下数据挖掘系统与数据库或数据仓库集成方法的差别：不耦合、松散耦合、半紧耦合和紧密耦合。你认为哪种方法最流行，为什么？

解答：

数据挖掘系统和数据库或数据仓库系统的集成的层次的差别如下。

- **不耦合：**数据挖掘系统用像平面文件这样的原始资料获得被挖掘的原始数据集，因为没有数据库系统或数据仓库系统的任何功能被作为处理过程的一部分执行。因此，这种构架是一种糟糕的设计。
- **松散耦合：**数据挖掘系统不与数据库或数据仓库集成，除了使用被挖掘的初始数据集的源数据和存储挖掘结果。这样，这种构架能得到数据库和数据仓库提供的灵活、高效、和特征的优点。但是，在大量的数据集中，由松散耦合得到高可测性和良好的性能是非常困难的，因为许多这种系统是基于内存的。
- **半紧密耦合：**一些数据挖掘原语，如聚合、分类、或统计功能的预计算，可在数据库或数据仓库系统有效的执行，以便数据挖掘系统在挖掘-查询过程的应用。另外，一些经常用到的中间挖掘结果能被预计算并存储到数据库或数据仓库系统中，从而增强了数据挖掘系统的性能。
- **紧密耦合：**数据库或数据仓库系统被完全整合成数据挖掘系统的一部份，并且因此提供了优化的数据查询处理。这样的话，数据挖掘子系统被视为一个信息系统的功能组件。这是一中高度期望的结构，因为它有利于数据挖掘功能、高系统性能和集成信息处理环境的有效实现。

从以上提供的体系结构的描述看，紧密耦合是最优的，没有值得顾虑的技术和执行问题。但紧密耦合系统所需的大量技术基础结构仍然在发展变化，其实现并非易事。因此，目前最流行的体系结构仍是半紧密耦合，因为它是松散耦合和紧密耦合的折中。

1.5 1.14 描述关于数据挖掘方法和用户交互问题的三个数据挖掘挑战。

第 2 章 数据预处理

2.1 2.2 假设给定的数据集的值已经分组为区间。区间和对应的频率如下。

年龄	频率
1~5	200
5~15	450
15~20	300
20~50	1500
50~80	700
80~110	44

计算数据的近似中位数值。

解答：

先判定中位数区间：N=200+450+300+1500+700+44=3194；N/2=1597

∵ 200+450+300=950<1597<2450=950+1500；

∴ 20~50 对应中位数区间。

我们有：L_1=20，N=3197，$(\sum freq)_l$=950，$freq_{median}$=1500，$width$=30，使用公式（2.3）：

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width = 20 + \left(\frac{3197/2 - 950}{1500} \right) \times 30 = 32.97$$

∴ $median$=32.97 岁。

2.2 2.4 假定用于分析的数据包含属性 age。数据元组的 age 值（以递增序）是：13，15，16，16，19，20，20，21，22，22，25，25，25，25，30，33，33，35，35，35，35，36，40，45，46，52，70。

(a) 该数据的均值是什么？中位数是什么？

(b) 该数据的众数是什么？讨论数据的峰（即双峰、三峰等）。

(c) 数据的中列数是什么？

(d) 你能（粗略地）找出数据的第一个四分位数（Q_1）和第三个四分位数（Q_3）

吗？

(e) 给出数据的五数概括。

(f) 画出数据的盒图。

(g) 分位数—分位数图与分位数图的不同之处是什么？

解答：

(a) 该数据的均值是什么？中位数是什么？

均值是：$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 809/27 = 29.96 \cong 30$（公式 2.1）。中位数应是第 14 个，即 $x_{14}=25=Q_2$。

(b) 该数据的众数是什么？讨论数据的峰（即双峰、三峰等）。

这个数集的众数有两个：25 和 35，发生在同样最高的频率处，因此是双峰众数。

(c) 数据的中列数是什么？

数据的中列数是最大术和最小是的均值。即：*midrange*=(70+13)/2=41.5。

(d) 你能（粗略地）找出数据的第一个四分位数（Q_1）和第三个四分位数（Q_3）吗？

数据集的第一个四分位数应发生在 25%处，即在(N+1)/4=7 处。所以：Q_1=20。而第三个四分位数应发生在 75%处，即在 3×(N+1)/4=21 处。所以：Q_3=35

(e) 给出数据的五数概括。

一个数据集的分布的 5 数概括由最小值、第一个四分位数、中位数、第三个四分位数、和最大值构成。它给出了分布形状良好的汇总，并且这些数据是：13、20、25、35、70。

(f) 画出数据的盒图。

略。

(g) 分位数—分位数图与分位数图的不同之处是什么？

分位数图是一种用来展示数据值低于或等于在一个单变量分布中独立的变量的粗略百分比。这样，他可以展示所有数的分位数信息，而为独立变量测得的值（纵轴）相对于它们的分位数（横轴）被描绘出来。

但分位数—分位数图用纵轴表示一种单变量分布的分位数，用横轴表示另一

单变量分布的分位数。两个坐标轴显示它们的测量值相应分布的值域，且点按照两种分布分位数值展示。一条线（y=x）可画到图中，以增加图像的信息。落在该线以上的点表示在y轴上显示的值的分布比x轴的相应的等同分位数对应的值的分布高。反之，对落在该线以下的点则低。

2.3 2.7 使用习题 2.4 给出的 age 数据回答下列问题：

(a) 使用分箱均值光滑对以上数据进行光滑，箱的深度为 3。解释你的步骤。评述对于给定的数据，该技术的效果。

(b) 如何确定数据中的离群点？

(c) 对于数据光滑，还有哪些其他方法？

解答：

(a) 使用分箱均值光滑对以上数据进行光滑，箱的深度为 3。解释你的步骤。评述对于给定的数据，该技术的效果。

用箱深度为 3 的分箱均值光滑对以上数据进行光滑需要以下步骤：

- 步骤 1：对数据排序。（因为数据已被排序，所以此时不需要该步骤。）
- 步骤 2：将数据划分到大小为 3 的等频箱中。

箱 1：13，15，16　　箱 2：16，19，20　　箱 3：20，21，22

箱 4：22，25，25　　箱 5：25，25，30　　箱 6：33，33，35

箱 7：35，35，35　　箱 8：36，40，45　　箱 9：46，52，70

- 步骤 3：计算每个等频箱的算数均值。
- 步骤 4：用各箱计算出的算数均值替换每箱中的每个值。

箱 1：44/3，44/3，44/3　箱 2：55/3，55/3，55/3　箱 3：21，21，21

箱 4：24，24，24　　箱 5：80/3，80/3，80/3　　箱 6：101/3，101/3，101/3

箱 7：35，35，35　　箱 8：121/3，121/3，121/3　箱 9：56，56，56

(b) 如何确定数据中的离群点？

聚类的方法可用来将相似的点分成组或“簇”，并检测离群点。落到簇的集外的值可以被视为离群点。作为选择，一种人机结合的检测可被采用，而计算机用一种事先决定的数据分布来区分可能的离群点。这些可能的离群点能被用人工轻松的检验，而不必检查整个数据集。

(c) 对于数据光滑，还有哪些其他方法？

其它可用来数据光滑的方法包括别的分箱光滑方法，如中位数光滑和箱边界光滑。作为选择，等宽箱可被用来执行任何分箱方式，其中每个箱中的数据范围均是常量。除了分箱方法外，可以使用回归技术拟合成函数来光滑数据，如通过线性或多线性回归。分类技术也能被用来对概念分层，这是通过将低级概念上卷到高级概念来光滑数据。

2.4 2.10 如下规范化方法的值域是什么？

(a) min-max 规范化。

(b) z-score 规范化。

(c) 小数定标规范化。

解答：

(a) min-max 规范化。

值域是[new_min, new_max]。

(b) z-score 规范化。

值域是[(old_min－mean)/σ，(old_max－mean)/σ]，总的来说，对于所有可能的数据集的值域是(－∞，+∞)。

(c) 小数定标规范化。

值域是(－1.0, 1.0)。

2.5 2.12 使用习题 2.4 给出的 age 数据，回答以下问题：

(a) 使用 min-max 规范化将 age 值 35 变换到[0.0，1.0]区间。

(b) 使用 z-score 规范化变换 age 值 35，其中 age 的标准差为 12.94 岁。

(c) 使用小数定标规范化变换 age 值 35。

(d) 对于给定的数据，你愿意使用哪种方法？陈述你的理由。

解答：

(a) 使用 min-max 规范化将 age 值 35 变换到[0.0，1.0]区间。

∵ $\min_A$=13，$\max_A$=70，$new\_\min_A$=0.0，$new\_\max_A$=1.0，而 v=35，

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

$$= \frac{35-13}{70-13}(1.0-0.0) + 0.0 = 0.3860$$

(b) 使用 z-score 规范化变换 age 值 35，其中 age 的标准差为 12.94 岁。

$$\overline{A} = \frac{13+15+2\times16+19+2\times20+21+2\times22+4\times25}{27} + \frac{30+2\times33+4\times35+36+40+45+46+52+70}{27} = \frac{809}{27} = 29.963$$

$$\sigma_A^2 = \frac{\sum_{i=1}^{N}(A_i-\overline{A})}{N} = 161.2949，\ \sigma_A = \sqrt{\sigma_A^2} = 12.7002$$

或 $$s_A^2 = \frac{\sum_{i=1}^{N}(A_i-\overline{A})}{N} = 167.4986，\ s_A = \sqrt{s_A^2} = 12.9421$$

v=35

$$v_\sigma' = \frac{v-\overline{A}}{\sigma_A} = \frac{35-29.963}{12.7002} = \frac{5.037}{12.7002} = 0.3966 \approx 0.400$$

或 $$v_s' = \frac{v-\overline{A}}{s_A} = \frac{35-29.963}{12.9421} = \frac{5.037}{12.9421} = 0.3892 \approx 0.39$$

(c) 使用小数定标规范化变换 age 值 35。

由于最大的绝对值为 70，所以 j=2。 $v' = \frac{v}{10^j} = \frac{35}{10^2} = 0.35$

(d) 对于给定的数据，你愿意使用哪种方法？陈述你的理由。

略。

2.6 2.14 假设 12 个销售价格记录组已经排序如下：5，10，11，13，15，35，50，55，72，92，204，215。使用如下每种方法将其划分成三个箱。

(a) 等频（等深）划分。

(b) 等宽划分。

(c) 聚类。

解答：

(a) 等频（等深）划分。

bin1	5,10,11,13
bin1	15,35,50,55

bin1	72,91,204,215

(b) 等宽划分。

每个区间的宽度是：(215-5)/3=70

bin1	5,10,11,13,15,35,50,55,72
bin1	91
bin1	204,215

(c) 聚类。

我们可以使用一种简单的聚类技术：用 2 个最大的间隙将数据分成 3 个箱。

bin1	5,10,11,13,15
bin1	35,50,55,72,91
bin1	204,215

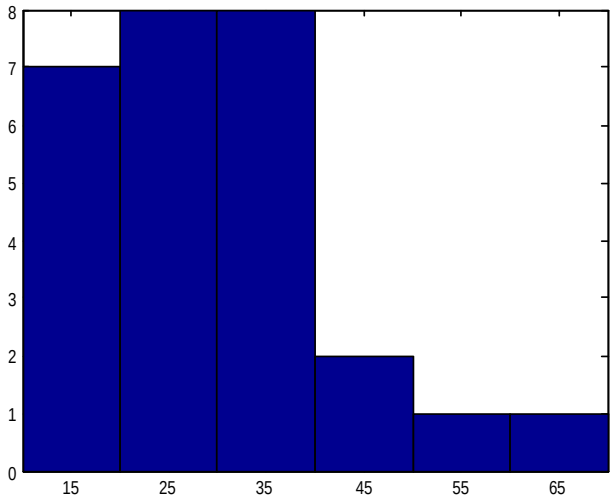
2.7 2.15 使用习题 2.4 给出的 age 数据，

(a) 画出一个等宽为 10 的等宽直方图；

(b) 为如下每种抽样技术勾画例子：SRSWOR，SRSWR，聚类抽样，分层抽样。使用大小为 5 的样本和层“青年”，“中年”和“老年”。

解答：

(a) 画出一个等宽为 10 的等宽直方图；



(b) 为如下每种抽样技术勾画例子：SRSWOR，SRSWR，聚类抽样，分层抽样。使用大小为 5 的样本和层“青年”，“中年”和“老年”。

元组：

T_1	13	T_{10}	22	T_{19}	35
T_2	15	T_{11}	25	T_{20}	35
T_3	16	T_{12}	25	T_{21}	35
T_4	16	T_{13}	25	T_{22}	36
T_5	19	T_{14}	25	T_{23}	40
T_6	20	T_{15}	30	T_{24}	45
T_7	20	T_{16}	33	T_{25}	46
T_8	21	T_{17}	33	T_{26}	52
T_9	22	T_{18}	35	T_{27}	70

SRSWOR 和 SRSWR：不是同次的随机抽样结果可以不同，但前者因无放回所以不能有相同的元组。

SRSWOR	(n=5)	SRSWR	(n=5)
T_4	16	T_7	20
T_6	20	T_7	20
T_{10}	22	T_{20}	35
T_{11}	25	T_{21}	35
T_{26}	52	T_{25}	46

聚类抽样：设起始聚类共有 6 类，可抽其中的 m 类。

Sample1		Sample2		Sample3		Sample4		Sample5		Sample6	
T_1	13	T_6	20	T_{11}	25	T_{16}	33	T_{21}	35	T_{26}	52
T_2	15	T_7	20	T_{12}	25	T_{17}	33	T_{22}	36	T_{27}	70
T_3	16	T_8	21	T_{13}	25	T_{18}	35	T_{23}	40		
T_4	16	T_9	22	T_{14}	25	T_{19}	35	T_{24}	45		
T_5	19	T_{10}	22	T_{15}	30	T_{20}	35	T_{25}	46		

Sample2	Sample5

T_6	20	T_{21}	35
T_7	20	T_{22}	36
T_8	21	T_{23}	40
T_9	22	T_{24}	45
T_{10}	22	T_{25}	46

分层抽样：按照年龄分层抽样时，不同的随机试验结果不同。

T_1	13	young	T_{10}	22	young	T_{19}	35	middle age
T_2	15	young	T_{11}	25	young	T_{20}	35	middle age
T_3	16	young	T_{12}	25	young	T_{21}	35	middle age
T_4	16	young	T_{13}	25	young	T_{22}	36	middle age
T_5	19	young	T_{14}	25	young	T_{23}	40	middle age
T_6	20	young	T_{15}	30	middle age	T_{24}	45	middle age
T_7	20	young	T_{16}	33	middle age	T_{25}	46	middle age
T_8	21	young	T_{17}	33	middle age	T_{26}	52	middle age
T_9	22	young	T_{18}	35	middle age	T_{27}	70	senior

T_4	16	young
T_{12}	25	young
T_{17}	33	middle age
T_{25}	46	middle age
T_{27}	70	Senior

2.8 55555555555555555555555555

第 3 章 数据仓库与 OLAP 技术概述

3.1 3.4 假定 BigUniversity 的数据仓库包含如下 4 个维：student(student_name, area_id, major, status, university)，course(course_name, department)，semester(semester, year)和 instructor(dept, rank)；2 个度量：count 和 avg_grade。在最低概念层，度量 avg_grade 存放学生的实际课程成绩。在较高概念层，avg_grade 存放给定组合的平均成绩。

(a) 为该数据仓库画出雪花形模式图。

(b) 由基本方体[student, course, semester, instructor]开始，为列出 BigUniversity 每个学生的 CS 课程的平均成绩，应当使用哪些特殊的 OLAP 操作。

(c) 如果每维有 5 层（包括 all），如“student<major<status<university<all”，该立方体包含多少方体？

解答：

a) 为该数据仓库画出雪花形模式图。雪花模式如图所示。

b) 由基本方体[student, course, semester, instructor]开始，为列出 BigUniversity 每个学生的 CS 课程的平均成绩，应当使用哪些特殊的 OLAP 操作。

这些特殊的联机分析处理（OLAP）操作有：

i. 沿课程（course）维从 course_id“上卷”到 department。

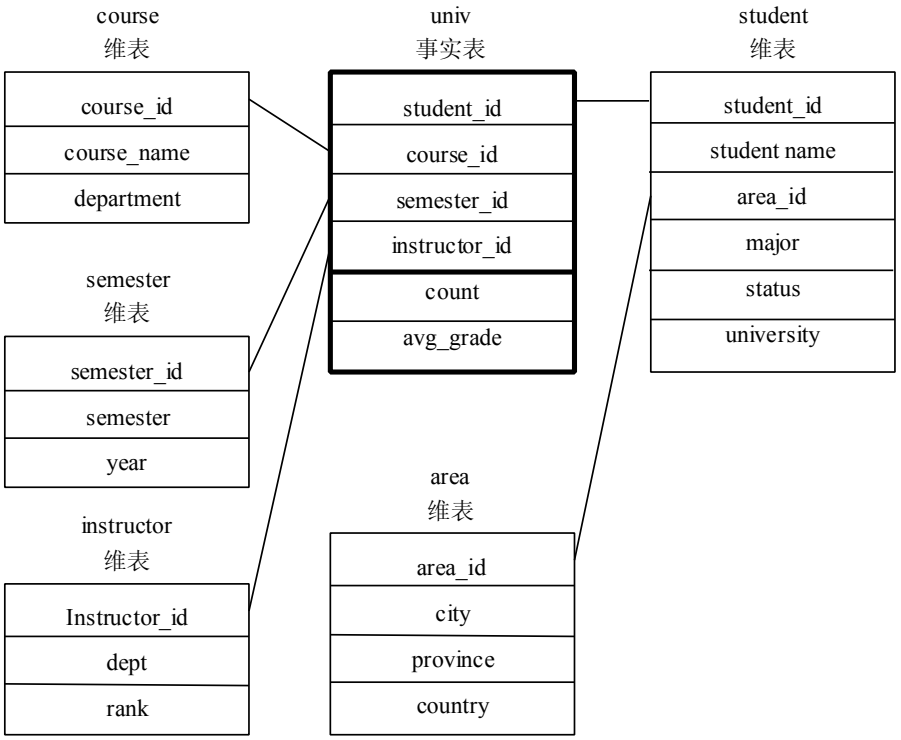
ii. 沿学生（student）维从 student_id“上卷”到 university。

iii. 取 department=“CS”和 university=“Big University”，沿课程（course）维和学生（student）维切片。

iv. 沿学生（student）维从 university 下钻到 student_name。

c) 如果每维有 5 层（包括 all），如“student<major<status<university<all”，该立方体包含多少方体？

这个立方体将包含 5^4=625 个方体。



题 3.4 图 题 3.4 中数据仓库的雪花形模式

3.2 2222222

3.3 3333333

第 4 章 数据立方体计算与数据泛化

4.1 2008-11-29

4.2 有几种典型的立方体计算方法，

4.3 题 4.12 考虑下面的多特征立方体查询：按{item，region，month}的所有子集分组，对每组找出 2004 年的最小货架寿命，并对价格低于 100 美元、货架寿命在最小货架寿命的 1.25~1.5 倍之间的元组找出总销售额部分。

d) 画出该查询的多特征立方体图。

e) 用扩充的 SQL 表示该查询。

f) 这是一个分布式多特征立方体吗？为什么？

解答：

(a) 画出该查询的多特征立方体图。

R0→R1(≥1.25*min(shelf)and≤1.5*min(shelf))

(b) 用扩充的 SQL 表示该查询。

```
select      item, region, month, Min(shelf), SUM(R1)
from        Purchase
where       year=2004
cube by     item, region, month: R1
such that   R1.shelf≥1.25*MIN(Shelf) and (R1.Shelf≤1.5*MIN(Shelf) and
R1.Price<100
```

(c) 这是一个分布式多特征立方体吗？为什么？

这不是一个分布多特征立方体，因为在“such that”语句中采用了“≤”条件。

4.4 2008-11-29

4.5 2008-11-29

第 5 章 挖掘频繁模式、关联和相关

5.1 Apriori 算法使用子集支持度性质的先验知识。

5.2 5.2.2 节介绍了由频繁项集产生关联规则的方法。提出了一个更有效的方法。解释它为什么比 5.2.2 节的方法更有效。（提示：考虑将习题 5.1(b)和习题 5.1(c)的性质结合到你的设计中。）

■

5.3 数据库有 5 个事物。设 min_sup=60%，min_conf=80。

TID	购买的商品
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

g) 分别使用 Apriori 和 FP 增长算法找出所有的频繁项集。比较两种挖掘过程的效率。

h) 列举所有与下面的的元规则匹配的强关联规则（给出支持度 s 和置信度 c），其中，X 是代表顾客的变量，item 是表示项的变量（如“A”、“B”等）：

$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$ [s, c]

解答：

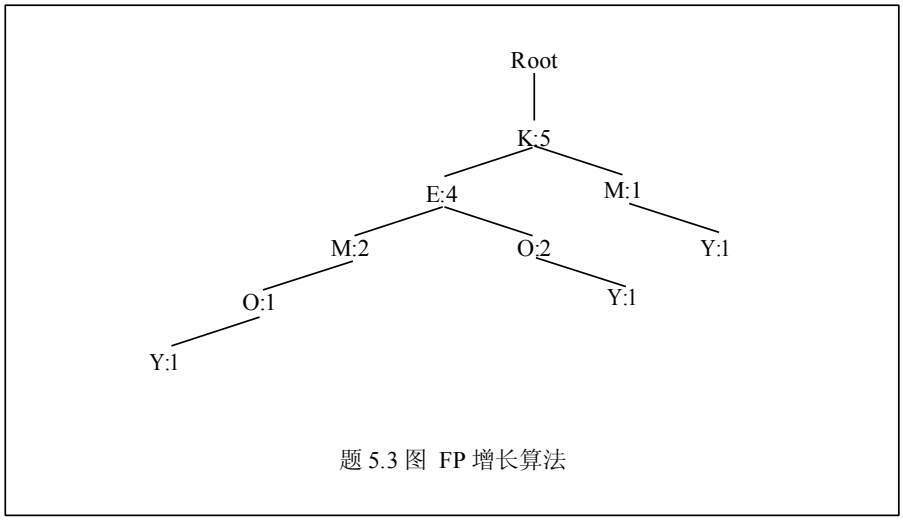
(a) 分别使用 Apriori 和 FP 增长算法找出所有的频繁项集。比较两种挖掘过程的效率。

Apriori 算法：由于只有 5 次购买事件，所以绝对支持度是 5×min_sup=3。

$$C_1=\begin{bmatrix} M & 3\\ O & 3\\ N & 2\\ K & 5\\ E & 4\\ Y & 3\\ D & 1\\ A & 1\\ U & 1\\ C & 2\\ I & 1 \end{bmatrix} \quad L_1=\begin{bmatrix} M & 3\\ O & 3\\ K & 5\\ E & 4\\ Y & 3 \end{bmatrix} \quad C_2=\begin{bmatrix} MO & 1\\ MK & 3\\ ME & 2\\ MY & 2\\ OK & 3\\ OE & 3\\ OY & 2\\ KE & 4\\ KY & 3\\ EY & 2 \end{bmatrix} \quad L_2=\begin{bmatrix} MK & 3\\ OK & 3\\ OE & 3\\ KE & 4\\ KY & 3 \end{bmatrix} \quad C_3=\begin{bmatrix} OKE & 3\\ KEY & 2 \end{bmatrix}$$

$$L_3=\begin{bmatrix} OKE & 3 \end{bmatrix}$$

FP-growth：数据库的第一次扫描与 Apriori 算法相同，得到 L_1。再按支持度计数的递减序排序，得到：L={(K:5), (E:4), (M:3), (O:3), (Y:3)}。扫描没个事务，按以上 L 的排序，从根节点开始，得到 FP-树。



题 5.3 图 FP 增长算法

项	条件模式基	条件 FP 树	产生的频繁模式
Y	{{K,E,M,O:1}，{K,E,O:1}，{K,M:1}}	K:3	{K,Y:3}
O	{{K,E,M:1}，{K,E:2}}	K:3，E:3	{K,O:3}，{E,O:3}，{K,E,O:3}
M	{{K,E:2}，{K:1}}	K:3	{K,M:3}
E	{{K:4}}	K:4	{K,E:4}

效率比较：Apriori 算法的计算过程必须对数据库作多次扫描，而 FP-增长算法在构造过程中只需扫描一次数据库，再加上初始时为确定支持度递减排序的一次扫描，共计只需两次扫描。由于在 Apriori 算法中的自身连接过程产生候选项集，候选项集产生的计算代价非常高，而 FP-增长算法不需产生任何候选项。

(b) 列举所有与下面的的元规则匹配的强关联规则（给出支持度 s 和置信度 c），其中，X 是代表顾客的变量，item 是表示项的变量（如“A”、“B”等）：

$\forall x \in$ transaction, buys(X, “K”) $\wedge$ buys(X, “O”) $\Rightarrow$ buys(X, “E”) [s=0.6, c=1]

$\forall x \in$ transaction, buys(X, “E”) $\wedge$ buys(X, “E”) $\Rightarrow$ buys(X, “K”) [s=0.6, c=1]

或也可表示为

K,O→E[s(support)=0.6 或 60%,c(confidence)=1 或 100%]

E,O→K[s(support)=0.6 或 60%,c(confidence)=1 或 100%]

■

5.4 （**实现项目**）使用你熟悉的程序设计语言（如 C++或 Java），实现本章介绍的三种频繁项集挖掘算法：

5.5 2008-12-01

5.6 2009-01-09

第 6 章 分类和预测

6.1 简述决策树分类的主要步骤。

6.2 6.11 下表由雇员数据库的训练数据组成。数据已泛化。例如，age“31…35”表示年龄在 31~35 之间。对于给定的行，count 表示 department，status，age 和 salary 在该行具有给定值的元组数。

department	status	age	salary	count
sales	senior	31…35	46K…50K	30
sales	junior	26…30	26K…30K	40
sales	junior	31…35	31K…35K	40
systems	junior	21…25	46K…50K	20
systems	senior	31…35	66K…70K	5
systems	junior	26…30	46K…50K	3
systems	senior	41…45	66K…70K	3
marketing	senior	36…40	46K…50K	10
marketing	junior	31…35	41K…45K	4
secretary	senior	46…50	36K…40K	4
secretary	junior	26…30	26K…30K	6

i) 如何修改基本决策树算法，以便考虑每个广义数据元组（即每一行）的 count？

j) 使用修改过的算法，构造给定数据的决策树。

k) 给定一个数据元组，它的属性 department，age 和 salary 的值分别为“systems”，“26…30”，和“46K…50K”。该元组 status 的朴素贝叶斯分类是什么？

l) 为给定的数据设计一个多层前馈神经网络。标记输入和输出层节点。

m) 使用上面得到的多层前馈神经网络，给定训练实例（sales，senior，31…35，46K…50K），给出后向传播算法一次迭代后的权重值。指出

你使用的初始权重和偏倚以及学习率。

解答：

(a) 如何修改基本决策树算法，以便考虑每个广义数据元组（即每一行）的 count？

(b) 使用修改过的算法，构造给定数据的决策树。

(c) 给定一个数据元组，它的属性 department，age 和 salary 的值分别为“systems”，“26…30”，和“46K…50K”。该元组 status 的朴素贝叶斯分类是什么？

解一：设元组的各个属性之间相互独立，所以先求每个属性的类条件概率：

P(systems|junior)=(20+3)/(40+40+20+3+4+6)=23/113;

P(26-30|junior)=(40+3+6)/113=49/113；

P(46K-50K|junior)=(20+3)/113=23/113；

∵ X=(department=system,age=26…30,salary=46K…50K)；

∴ P(X|junior)=P(systems|junior)P(26-30|junior)P(46K-50K|junior)

$=23\times49\times23/113^3=25921/1442897=0.01796$；

P(systems|senior)=(5+3)/(30+5+3+10+4)=23/52;

P(26-30|senior)=(0)/53=0;

P(46K-50K|senior)=(30+10)/52=40/52；

∵ X=(department=system,age=26…30,salary=46K…50K)；

∴ P(X|senior)=P(systems|senior)P(26-30|senior)P(46K-50K|senior)=0;

∵ P(junior)=113/165=0.68；

∵ P(senior)=52/165=0.32；

∴ P(X|junior)P(junior)=0.01796×0. 68=0. 0122128>0=0=P(X|senior)P(senior)；

所以：朴素贝叶斯分类器将 X 分到 junior 类。

解二：设元组的各属性之间不独立，其联合概率不能写成份量相乘的形式。

所以已知：X=(department=system,age=26…30,salary=46K…50K)，元组总数为：30+40+40+20+5+3+3+10+4+4+6=165。

先验概率：

当 status=senior 时，元组总数为：30+5+3+10+4=52，P(senior)=52/165=0.32；

当 status=junior 时，元组总数为： 40+40+20+3+4+6=113，P(junior)=113/165=0.68；

因为 status=senior 状态没有对应的 age=26…30 区间，所以：P(X|senior)=0；

因为 status=junior 状态对应的 partment=systems、age=26…30 区间的总元组数为：3，所以：P(X|junior)=3/113；

因为：P(X|junior)P(junior)=3/113×113/165＝0.018>0=P(X|senior)P(senior)；

所以：朴素贝叶斯分类器将 X 分到 junior 类。

(d) 为给定的数据设计一个多层前馈神经网络。标记输入和输出层节点。

(e) 使用上面得到的多层前馈神经网络，给定训练实例（sales，senior，31…35，46K…50K），给出后向传播算法一次迭代后的权重值。指出你使用的初始权重和偏倚以及学习率。

6.3 2008-12-01

6.4 2008-12-01

第 7 章 聚类分析

第 8 章 流挖掘、时间序列和序列数据

第 9 章 图挖掘、社会网络分析和多关系数据挖掘

第 10 章 挖掘对象、空间、多媒体、文本和 Web 数据

第 11 章数据挖掘的应用和发展趋势