2015 年数据仓库与知识发现试题

1. 数据仓库及其实现技术。

   a) 试采用 BITMAP 索引方式对图 1 中的维度表进行索引。

   b) 试采用 Join Index 对图 1 中的事实表和维表进行索引。

PID	SKU	TYPE	PRICE
01	BK-6573	BOOK	High
02	SW-8761	SOFTWARE	High
03	BK-7651	BOOK	Middle
04	CD-3413	CD	Middle
05	CD-6573	CD	Free
06	SW-9871	SOFTWARE	Middle

SID	Manager	TYPE
01	Bob	General
02	John	Exclusive
03	Smith	General

PID	SID	TID	Quantity
03	03	T100	3
01	01	T200	7
04	02	T300	5
02	03	T400	1
04	02	T500	2
05	01	T600	4
01	03	T700	6
05	02	T900	1
06	01	T900	6
01	02	T1000	3

图 1 产品维度表（左上）、商店维度表（左下）和销售事实表（右

2. 特征

给定图 2 中的目标集（DOG）和对比集（CAT），使用信息增益计算各个属性与当前概念描述任务之间的相关性。并采用 T=0.1 作为阈值，对属性进行筛选。

Gender	Tail	Weight	Count
M	Long	5-10	2
M	Middle	0-5	3
F	Long	5-10	3
M	Middle	10-15	1
M	Short	10-15	3
F	Long	15-20	3

Gender	Tail	Weight	Count
M	Long	0-5	2
F	Middle	5-10	1
F	Short	0-5	2
F	Long	5-10	2
M	Middle	0-5	1
F	Short	5-10	2

图 2 目标集 DOG（左）、对比集 CAT（右）

3. 关联

a) 针对图 3 的交易事务数据，采用 FP 增长算法求取频繁项集，假设最小支持度为≥30%

事务 ID	购买项
1	{a，b，d，e}
2	{b，c，d}
3	{a，b，d，e}
4	{a，c，d，e}
5	{b，c，d，e}
6	{b，d，e}
7	{c，d}
8	{a，b，c}
9	{a，d，e}
10	{b，d}

图 3 交易事务数据

基于上述频繁项集，构造关联规则，要求最小置信度≥50%

4. 数据预处理与分类(25分)

a) 针对图4中训练数据集进行离散化处理。要求采用等宽分桶的方式将age和incoming属性离散到3个区间。

b) 依据训练集，采用朴素贝叶斯方法分类未知元组(24, 75000, yes)，对分类属性Class:buys_MP进行预测。

ID	age	income	student	Class:buys_MP
1	23	68000	no	>2000
2	49	36000	no	1000..2000
3	55	22000	no	1000..2000
4	34	30000	yes	<1000
5	38	15000	yes	<1000
6	57	75000	no	>2000
7	21	52000	no	1000..2000
8	31	45000	yes	1000..2000
9	66	58000	no	1000..2000
10	34	12000	yes	<1000
11	40	40000	yes	1000..2000
12	50	78000	no	>2000
13	29	20000	yes	1000..2000
14	25	70000	no	<1000
15	61	55000	no	>2000
16	45	65000	no	>2000

图4 训练数据集

5. 聚类（25分）

a) 针对图5中的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵。

b) 采用凝聚式层次式方法对该数据集进行聚类，聚类间的距离使用聚类中数据之间的最大距离进行度量。

ID	x	y
1	3	8
2	2	7
3	4	8
4	3	4
5	4	5

图5 聚类数