

## 2012 年数据仓库与知识发现试题

### 1. 数据仓库及其实现技术。(25 分)

- 简述数据仓库在知识发现过程中的作用和地位。
- 为何 B 树等在数据库中广泛使用的索引技术无法被直接引入数据仓库?
- 试采用 BITMAP 索引方式对图 1 中的维度表进行索引。

ID	SKU	TYPE	PRICE
01	BK-6573	BOOK	High
02	CD-7189	CD	Low
03	SW-8761	SOFTWARE	High
04	BK-7651	BOOK	Middle
05	CD-3413	CD	Middle
06	BK-9861	BOOK	Free
07	CD-6573	CD	Free
08	SW-9871	SOFTWARE	Middle
09	CD-7123	CD	Low
10	BK-7123	BOOK	High

图 1 产品维度表

### 2. 关联 (25 分)

图 1 产品维度表

### 2. 关联 (25 分)

- 针对图 2 的交易事务数据, 采用 Apriori 算法求取频繁项集, 假设最小支持度为  $\geq 30\%$

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 2 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  $\geq 50\%$



### 3. 数据预处理与分类(25分)

- a) 针对图3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, 采用信息增益作为指标构造决策树。
- c) 采用构造出的决策树, 分类未知元组 (24, 75000, yes)。

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3 训练数据集

16	45	65000	no	>2000
----	----	-------	----	-------

图3 训练数据集

### 4. 聚类(25分)

- a) 针对下图的数据, 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 K-平均点方法对该数据集进行聚类, 其中  $K=3$ , 起始中心点 ID=1, ID=2, ID=3, 即, (3, 5); (2, 6); (3, 8)。

ID	x	y
1	3	5
2	2	6
3	3	8
4	3	4
5	7	7
6	4	5
7	9	1
8	4	10
9	1	6
10	6	8
11	5	2
12	4	2

图4 聚类数据集



age:  $d = \frac{66 - 21}{3} = 15$

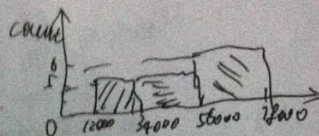
(21, 26], (36, 51], (51, 66]

① ↑      ② ↑      ③ ↑

income:  $d = \frac{78000 - 12000}{3} = 22000$

[12000, 34000], (34000, 56000], (56000, 78000]

③ ↑      ⑤ ↑      ⑥ ↑



(2) ① ② ③ class: buys-mp < 1000 4

④ ⑤ ⑥ class: buys-mp 1000...2000 7

⑦ ⑧ ⑨ class: buys-mp > 2000 5

$$I(s_1, s_2, s_3) = I(4, 7, 5) = -\frac{4}{16} \log_2 \frac{4}{16} - \frac{7}{16} \log_2 \frac{7}{16} - \frac{5}{16} \log_2 \frac{5}{16} = 1.544$$

age = "a [21, 36]  $s_{11}=2$   $s_{21}=3$   $s_{31}=1$

[36, 51]  $s_{12}=1$   $s_{22}=2$   $s_{32}=2$

[51, 66]  $s_{13}=0$   $s_{23}=2$   $s_{33}=2$

$4+7+5 = 16$

$I(s_{11}, s_{21}, s_{31}) = 1.45$

$I(s_{12}, s_{22}, s_{32}) = 1.522$

$I(s_{13}, s_{23}, s_{33}) = 1$

$$E(\text{age}) = \frac{7}{16} I(s_{11}, s_{21}, s_{31}) + \frac{5}{16} I(s_{12}, s_{22}, s_{32}) + \frac{4}{16} I(s_{13}, s_{23}, s_{33}) = 1.360$$

$$\text{Gain}(\text{age}) = I(s_1, s_2, s_3) - E(\text{age}) = 0.18$$

income

[12000, 34000]  $s_{11}=3$   $s_{21}=2$   $s_{31}=0$   $I(s_{11}, s_{21}, s_{31}) = 0.9710$

[34000, 56000]  $s_{12}=0$   $s_{22}=4$   $s_{32}=1$   $I(s_{12}, s_{22}, s_{32}) = 0.7219$

[56000, 78000]  $s_{13}=1$   $s_{23}=1$   $s_{33}=3$   $I(s_{13}, s_{23}, s_{33}) = 0.8422$

$$E(\text{income}) = \frac{5}{16} \times 0.9710 + \frac{5}{16} \times 0.7219 + \frac{6}{16} \times 0.8422 = 0.8449$$

$$\text{Gain}(\text{income}) = I - E = 0.6991$$



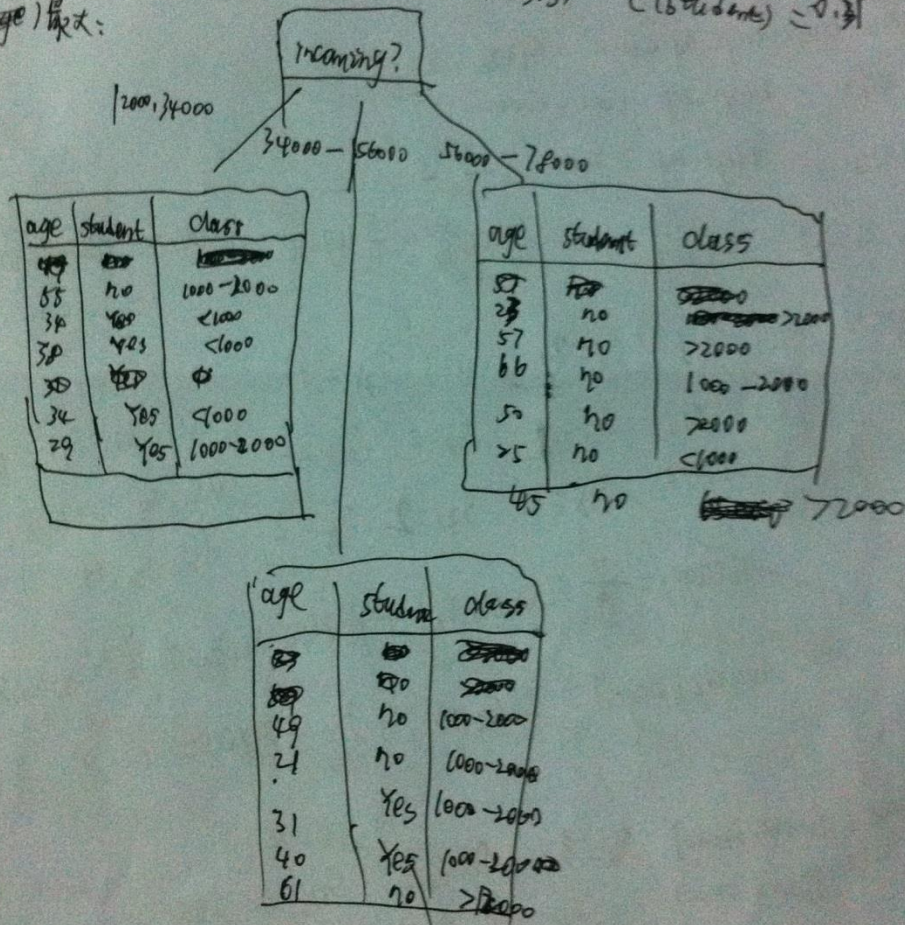
student yes  $S_{11}=3$   $S_{21}=3$   $S_{31}=0$   $I(S_{11}, S_{21}, S_{31})=1$

no  $S_{12}=1$   $S_{22}=4$   $S_{32}=5$   $I(S_{12}, S_{22}, S_{32})=1.3610$

$$E(\text{student}) = \frac{6}{16} \times I(S_{11}, S_{21}, S_{31}) + \frac{10}{16} \times I(S_{12}, S_{22}, S_{32}) = 1.2256$$

incoming  
E(age)最大:

$$\text{Gain}(\text{student}) = I(S_1, S_2, S_3) - E(\text{student}) = 0.37$$





例 1:

例 2:

Yes  
No  
Yes

age	student	class
35	no	1000-2000
34	yes	<1000
38	yes	<1000
34	yes	<1000
29	yes	1000-2000

Income

age	student	class
23	no	>2000
57	no	>2000
66	no	1000-2000
50	no	>2000
25	no	>2000
45	no	>2000

age

age	student	class
49	no	1000-2000
21	no	1000-2000
31	yes	1000-2000
40	yes	1000-2000
61	no	>2000

age

S	C
no	>2000
no	>2000

S	C
no	>2000
no	>2000

S	C
no	>2000
yes	>2000

1000-2000

$$S_1 = 3, S_2 = 2$$

$$I(S_1, S_2) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.9710$$

$$\text{对于 age } [2, 36] \quad S_{11} = 2 \quad S_{21} = 1$$

$$\text{age } [36, 51] \quad S_{12} = 1 \quad S_{22} = 0$$

$$\text{age } [51, 66] \quad S_{13} = 0 \quad S_{23} = 1$$

$$I(S_{11}, S_{21}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$I(S_{12}, S_{22}) = -\frac{1}{1} \log_2 \frac{1}{1} \times 0 = 0$$

$$I(S_{13}, S_{23}) = -0 \times -\frac{1}{1} \log_2 1 = 0$$

$$E(\text{age}) = \frac{2}{5} \times 0.9183 = 0.4592$$

$$\text{Gain}(\text{age}) = I(S_1, S_2) - E(\text{age}) = 0.4510$$

对于 student

$$\begin{matrix} \text{yes} & S_{11} = 3 & S_{21} = 1 \\ \text{no} & S_{12} = 0 & S_{22} = 1 \end{matrix}$$

$$I(S_{11}, S_{21}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

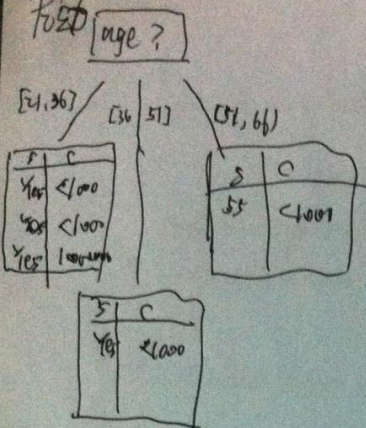
$$I(S_{12}, S_{22}) = 0 - \frac{1}{1} \log_2 1 = 0$$

$$E(\text{student}) = \frac{3}{4} \times 0.8113 = 0.6092$$

$$\text{Gain}_2 = I(S_1, S_2) - E(\text{student}) = 0.33 \quad \text{按 age 分}$$



$\frac{4}{5} = \frac{4}{5}$   
first



~~age~~

$$I(S_1, S_2) = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} = 0.7219$$

for Age [21, 36]  $S_{11}=2, S_{21}=0, I(S_{11}, S_{21})=0$

[36, 51]  $S_{12}=2, S_{22}=0, I(S_{12}, S_{22})=0$

[51, 66]  $S_{13}=0, S_{23}=1, I(S_{13}, S_{23})=0$

for student  $E(\text{age}) = 0, G(\text{cash}) = 0.7219$

Yes  $S_{11}=2, S_{21}=0, I(S_{11}, S_{21})=0$

No  $S_{12}=2, S_{22}=1, I(S_{12}, S_{22}) \neq 0 > 0$

$$E(\text{student}) = \frac{2}{5} \times I(S_{12}, S_{21}) > 0$$

$$G(\text{student}) < G(\text{cash})$$

age is

$\frac{1}{5} = \frac{1}{5}$

