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）

   a) 简述数据仓库在知识发现过程中的作用和地位。

   b) 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？

   c) 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）

   a) 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≥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

   b) 基于上述频繁项集，构造关联规则，要求最小置信度≥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

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

4. 聚类（25分）

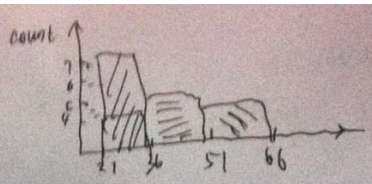
a) 针对下图的数据，采用曼哈顿距离作为距离函数，给出对应的相异矩阵。

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)。

(1) age: $d = \frac{\text{max} - \text{min}}{3} = \frac{66-21}{3} = 15$

[21, 36], (36, 51], (51, 66]

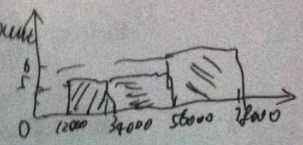
①个 ⑤个 ④个



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

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

③个 ⑧个 ⑤个



(2) 类标签:

C1 对应 class: buys_mp ≤ 1000 有4个

C2 对应: buys_mp 1000…2000 … 7

C3 对应 buys_mp > 2000 … 5

$I(S_1, S_2, S_3) = I(4, 7, 5) = -\frac{4}{16}\log_2\frac{4}{16} - \frac{7}{16}\log_2\frac{7}{16} - \frac{5}{16}\log_2\frac{5}{16} =$

$= 1.544$

对age = 区间 [21, 36] $S_{11}=3$ $S_{21}=3$ $S_{31}=1$ 4+7+5.

[36, 51] $S_{12}=1$ $S_{22}=2$ $S_{32}=2$ $I(S_{11}, S_{21}, S_{31}) = 1.45$.

[51, 66] $S_{13}=0$ $S_{32}=2$ $S_{33}=2$. $I(S_{12}, S_{22}, S_{32}) = 1.522$.

$I(S_{13}, S_{32}, S_{33}) = 1$.

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

$Gain(age) = I(S_1, S_2, S_3) - E(age) = 0.18$.

对于 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] $S_{13}=1$ $S_{32}=1$ $S_{33}=3$ $I(S_{13}, S_{32}, S_{33}) = 0.8422$

$E(\text{income}) = \frac{5}{16} \times 0.9710 + \frac{5}{16} \times 0.7219 + \frac{6}{16} \times 0.8422 = 0.84469$

$Gain(\text{income}) = 1.544 - 0.8449 = 0.6991$

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

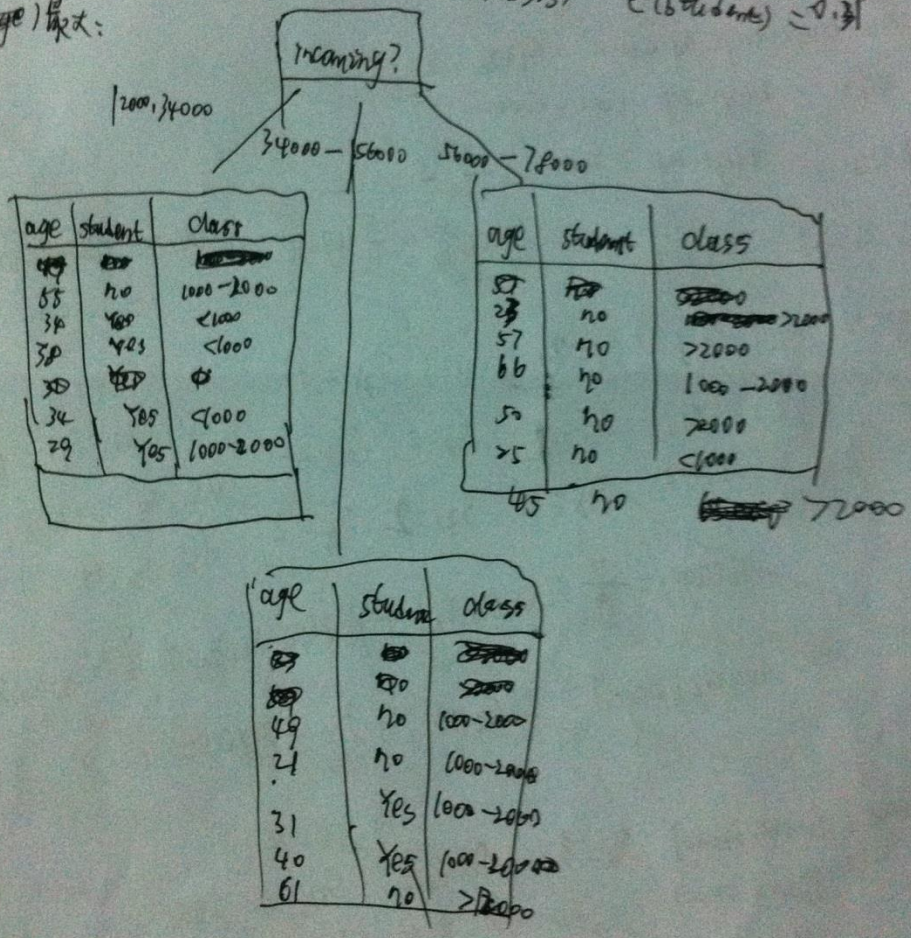
no $S_{12}=1$ $S_{22}=4$ $S_{32}=5$ $I(S_{12},S_{22},S_{32})=1.3610$

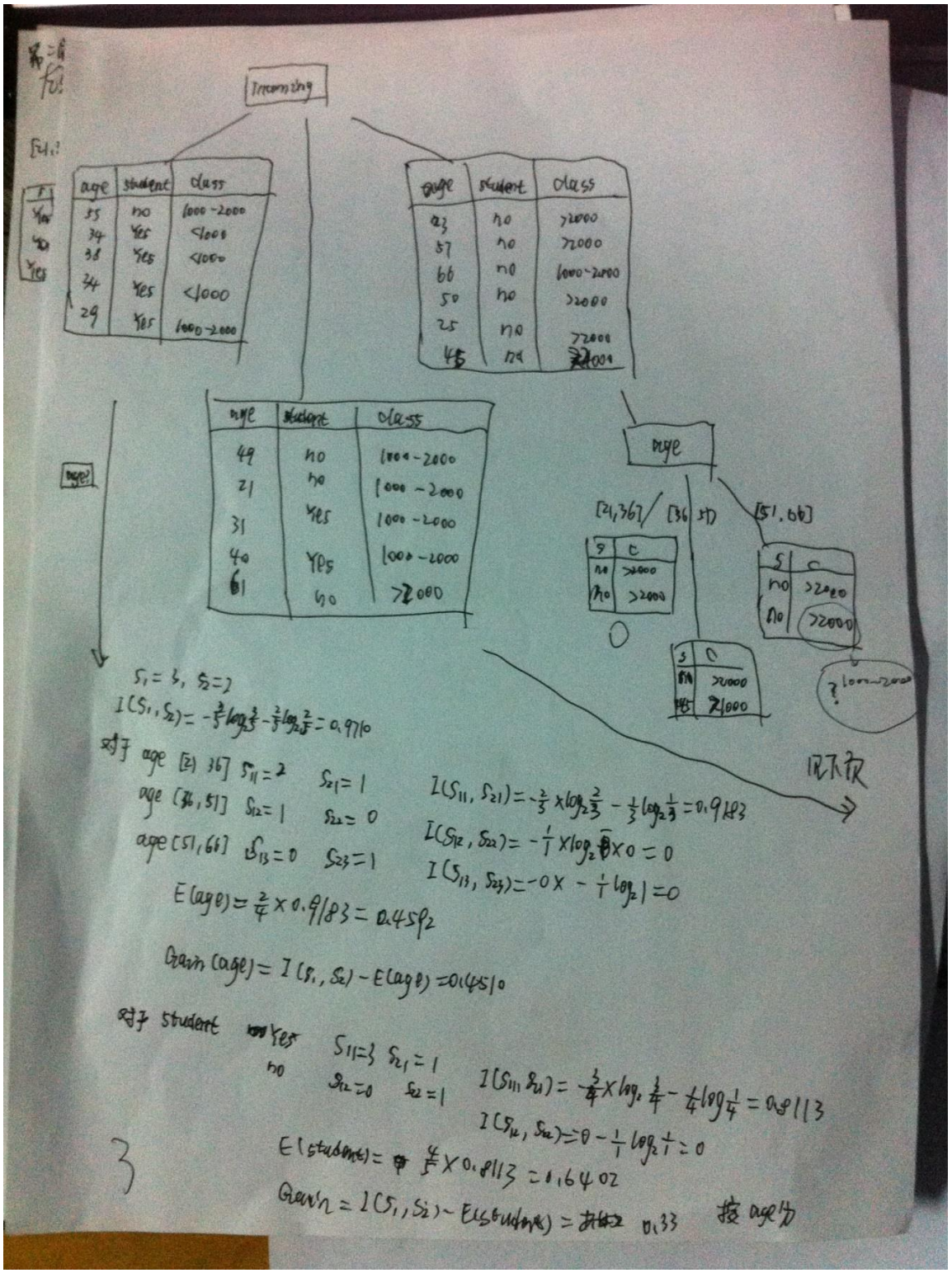
$E(\text{student})=\frac{6}{16}\times I(S_{11},S_{21},S_{31})+\frac{10}{16}\times I(S_{12},S_{22},S_{32})=1.2256$

Gain(student) $=I(S_1,S_2,S_3)-E(\text{student})=0.3$

incoming

E(age)最大：




Income/log
age | student | class
student
class
age
[21,36] [36,51] [51,66]
?
S1 = 5, S2 = 2
I(S1, S2) = -2/5 log 2/5 - 2/5 log 2/5 = 0.970
对于 age [21,36] S11 = 2 S21 = 1
age [36,51] S12 = 1 S22 = 0
age [51,66] S13 = 0 S23 = 1
I(S11, S21) = -2/3 x log2 2/3 - 1/3 log2 1/3 = 0.918
I(S12, S22) = -1/1 x log2 1 x 0 = 0
I(S13, S23) = -0 x - 1/1 log2 1 = 0
E(age) = 2/4 x 0.918 = 0.459
Gain(age) = I(S1, S2) - E(age) = 0.511
对于 student yes S11 = 3 S21 = 1
no S12 = 0 S22 = 1
I(S11, S21) = -3/4 x log2 3/4 - 1/4 log2 1/4 = 0.811
I(S12, S22) = 0 - 1/1 log2 1/1 = 0
E(student) = 4/5 x 0.811 = 0.6402
Gain = I(S1, S2) - E(student) = 0.33
按 age 分

age?

[21,36] [36,51] [51,66]

S	C
Yes	<1000
No	<1000
Yes	1000~2000

S	C
No	<1000

S	C
Yes	>2000

$I(S_1, S_2) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72$

对于 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] $S_{13}=0$ $S_{23}=1$ $I(S_{13}, S_{23})=0$

$E(age) = 0$, $Gain(age) = 0.72$

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

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

$E(student) = \frac{2}{5} \times I(S_{12}, S_{22}) > 0$

$Gain(student) < Gain(age)$

按 age 分

age?

[21,36] [36,51] [51,66]

S	C
Yes	1000~2000
Yes	1000~2000

S	C
No	2000

S	C
No	1000~2000
Yes	1000~2000

4

相

1	0									
2	2	0								
3	3	3	0							
4	1	3	4	0						
5	6	6	5	7	0					
6	1	3	4	2	5	0				
7	10	12	13	9	8	9	0			
8	6	6	3	7	6	5	14	0		
9	3	1	4	4	7	4	13	7	0	
10	6	6	3	7	2	5	10	4	7	0
11	5	7	8	4	7	4	5	9	8	7	0
12	4	6	7	3	8	3	6	8	7	8	1

(2) ID=1 (3,5) (2,6) , (3,8)

第一次循环: ~~(3,5)~~ (3,5) (3,4) (4,5) (9,1) (5,2) (4,2) K1(4.6667, 3.1667)

(2,6) (1,6) K2(1.5, 6.0)

(3,8) (7,7) (4,10) (6,8) K3(5.0, 8.25)

二:

(3,4) (4,5) (9,1) (5,2) (4,2) K1(5.0, 2.8)

(3,5) (2,6) (1,6) K2(2, 5.6667)

(3,8) (7,7) (4,10) (6,8) K3(5.0, 8.25)

三: (9,1) (5,2) (4,2) K1(6.0, 1.6667)

(3,5) (2,6) (3,4) (4,5) (1,6) K2(2.6, 5.2)

(3,8) (7,7) (4,10) (6,8) (5, 8.25)

三次循环