数据挖掘第二次作业

BraveY

2020 年 1 月 17 日

1 第二次作业

1.1 题目1

1. Consider the data set shown in Table 1（$min\_sup = 60\%, min\_conf = 70\%$）.

(a) Find all frequent itemsets using Apriori by treating each transaction ID as a market basket.

(b) Use the results in part (a) to compute the confidence for the association rules $\{a, b\} \rightarrow \{c\}$ and $\{c\} \rightarrow \{a, b\}$. Is confidence a symmetric measure?

(c) List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g. “A”, “B”, etc.):

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

TID	Items-bought
T1	A,D,B,C
T2	D,A,C,E,B
T3	A,B,E
T4	A,B,D

表 1: Example of market basket transactions

1.1.1 解

(a) 依题意最小的支持度计数为$sup = 4 \times 0.6 = 2.4$,首先计算所有频繁1项集的候选C_1如表2所示,候选的支持度计数与最小支持度计数比较得到频繁1项集如表3所示。根据L1生成候选2项集如4所示,候选的支持度计数与最小支持度计数比较得到频繁1项集如表5所示。根据L2生成候选3项集如6所示,候选的支持度计数与最小支持度计数比较得到频繁1项集如表7所示。所以所有的频繁项集如表8所示。

itemsets	sup
A	4
B	4
C	2
D	3
E	2

表 2: C1

itemsets	sup
A	4
B	4
D	3

表 3: L1

itemsets	sup
A	4
B	4
D	3

表 4: C2

(b) $\{a, b\} \rightarrow c$的置信度为:

$$confidence(\{a, b\} \rightarrow c) = \frac{support\_count(\{a, b, c\})}{support\_count(a, b)} = \frac{2}{4} = 50\%$$

itemsets	sup
A,B	4
A,D	3
B,D	3

表 5: L2

itemsets	sup
A,B,D	3

表 6: C3

$c \to \{a, b\}$的置信度为：

$$confidence(c \to \{a, b\}) = \frac{support\_count(a.b, c)}{support\_count(\{c\})} = \frac{2}{2} = 100\%$$

可以看出置信度不是一个对称的指标。

(c)根据a小题得到的频繁3项集为：$L = \{A, B, D\}$,其支持度为3。L的所有非空子集有$\{A, B\}, \{A, D\}, \{B, D\}, \{A\}, \{B\}, \{D\}$,因此所有能产生的符合要求的关联规则为：

$$\begin{aligned} \{A, B\} &\to \{D\}[s = \frac{3}{4} = 75\%, c = \frac{3}{4} = 75\%] \\ \{A, D\} &\to \{B\}[s = \frac{3}{4} = 75\%, c = \frac{3}{3} = 100\%] \\ \{B, D\} &\to \{A\}[s = \frac{3}{4} = 75\%, c = \frac{3}{3} = 100\%] \end{aligned} \tag{1}$$

1.2 题目2

2. Consider the data set shown in Table 1（$min_sup = 60\%$）.

(a) Find all frequent itemsets using FP-Growth. Please present all the FP-trees and all the conditional pattern bases.

(b) Compare the efficiency of Apriori and FP-Growth.

1.2.1 解

(a)所有的FP树如图 1所示。所有的条件模式基如表9所示。

(b)Apriori算法因为当频繁1项集很大时会产生大量的候选项集，而且每次增加频繁项集的大小，Apriori算法都会重新扫描整个数据集。数据集

itemsets	sup
A,B,D	3

表 7: L3

itemsets	sup
A	4
B	4
D	3
A,B	4
A,D	3
B,D	3
A,B,D	3

表 8: total frequent itemsets

很大时，这会显著降低频繁项集发现的速度。FP-growth算法基于Apriori构建，但采用了高级的数据结构减少扫描次数，大大加快了算法速度。

1.3 题目3

3. Suppose that the data mining task is to cluster the following ten points (with(x, y, z) representing location) into three clusters: A1(4,2,5), A2(10,5,2), A3(5,8,7), B1(1,1,1), B2(2,3,2), B3(3,6,9), C1(11,9,2),C2(1,4,6), C3(9,1,7), C4(5,6,7)

The distance function is Euclidean distance. Suppose initially we assign A1,B1,C1 as the center of each cluster, respectively. Use the K-Means algorithm to show only

(a) The three cluster center after the first round execution

(b) The final three clusters

1.3.1 解

(a)第一轮迭代中有每个点到中心点的距离如表10所示。其中C_i表示到第ω_i类的中心，在第一轮有$C_1 = A1, C_2 = B1, C_3 = C1$。 (b)第一轮迭代

Item	conditional pattern	frequent patterns
D	{A,B:3}	{A,B:3},{B,D:3},{A,B,D:3}
B	{A:4}	{A,B:4}

表 9: 条件模式基

点	到C_1的距离	到C_2距离	到C_3的距离	属于
A1	0	5.10	10.34	ω_1
A2	7.34	9.90	4.12	ω_3
A3	6.40	10.05	7.87	ω_1
B1	5.10	0	12.85	ω_2
B2	3.74	2.45	10.82	ω_2
B3	5.75	9.64	11.05	ω_1
C1	10.34	12.85	0	ω_3
C2	3.74	5.83	11.87	ω_1
C3	5.48	10	9.64	ω_1
C4	4.58	8.77	8.37	ω_1

表 10: K-Means第一轮迭代

后，形成的新的中心点为：

$$C_1 = (4.5, 4.5, 6.83), C_2 = (1.5, 2, 1.5), C_3 = (10.5, 7, 2)$$

第二轮迭代如表10所示。其中C_i表示到第ω_i类的中心，在第一轮有$C_1 = A1, C_2 = B1, C_3 = C1$。 因为与第一轮迭代的分类结果相同，所以不需要再进行迭代，第二轮就是最后一轮迭代。

1.4 题目4

User-Product rating matrix is shown in table 12:

(a) List the top 3 most similar users of user 2 based on Cosine Similarity.

(b) Predict User 2’ s rating for Product 2.

1.4.1 解

(a) 计算出其余用户与用户2的余弦相似度如表13所示。

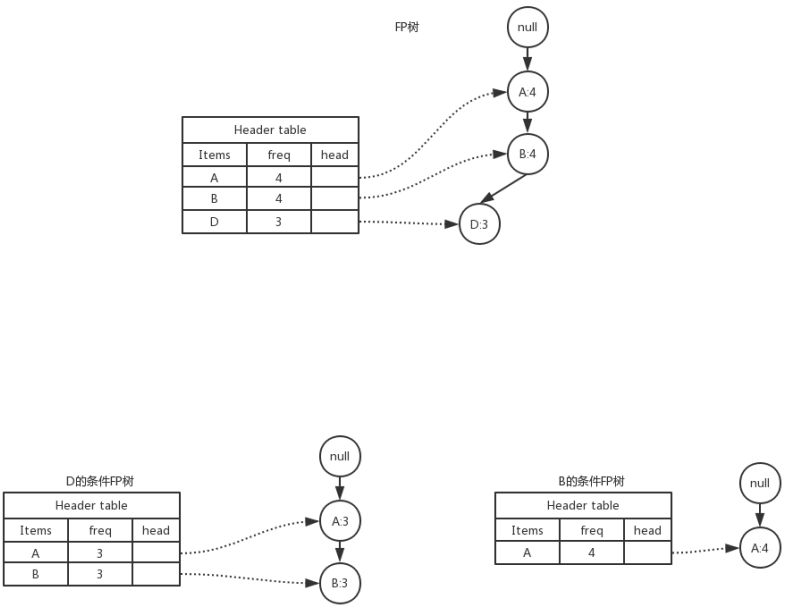


图 1: FP-tree

（b）对与用户2最相似的3个用户:用户3、用户1、用户5计算平均的打分表如表14所示。

因此可以计算出：

$$
\begin{aligned}
r_{User2,Product2} = \bar{r}_{User2} &+ \frac{sim(User2,User3)(r_{User3,Product2} - \bar{r}_{User3})}{sim(User2,User3) + sim(User2,User1) + sim(User2,User5)} \\
&+ \frac{sim(User1,User3)(r_{User1,Product2} - \bar{r}_{User1})}{sim(User2,User3) + sim(User2,User1) + sim(User2,User5)} \\
&+ \frac{sim(User2,User5)(r_{User5,Product2} - \bar{r}_{User5})}{sim(User2,User3) + sim(User2,User1) + sim(User2,User5)} \\
&= 4 + \frac{0.98 \times (3 - 1.5) + 0.96 \times (1 - 2.5) + 0.92 \times (2 - 2.5)}{0.98 + 0.96 + 0.92} \\
&= 3.85
\end{aligned}
$$

所以预测用户2对商品2的打分为3.85分。

点	到C_1的距离	到C_2距离	到C_3的距离	属于
A1	3.14	4.30	8.73	ω_1
A2	7.34	9.03	2.06	ω_3
A3	3.54	8.86	7.5	ω_1
B1	7.65	1.22	11.28	ω_2
B2	5.64	1.22	9.39	ω_2
B3	3.03	8.63	10.30	ω_1
C1	9.26	11.81	2.06	ω_3
C2	3.63	4.95	10.74	ω_1
C3	5.70	9.35	7.95	ω_1
C4	1.59	7.64	7.5	ω_1

表 11: K-Means第二轮迭代

	Product1	Product2	Product3	Product4
User1	1	1	5	3
User2	3	?	5	4
User3	1	3	1	1
User4	4	3	2	1
User5	2	2	2	4

表 12: User-Product Rating Matrix

1.5 Lab Part

1.5.1 Question1

按照提升度排序的结果如图2所示，其中前5条规则为：

$$\begin{aligned} &\{tomato\ source\} \rightarrow \{pasta\} \\ &\{coffee, milk\} \rightarrow \{pasta\} \\ &\{biscuits, pasta\} \rightarrow \{milk\} \\ &\{pasta, water\} \rightarrow \{milk\} \\ &\{juices\} \rightarrow \{milk\} \end{aligned} \tag{2}$$

因为上面5条规则都是满足提升度大于1，所以是正相关的，所以这5条规则都是有趣的强关联规则。上面的规则没有冗余的，因此可以将tomato

	与用户2的余弦相似度
User3	0.98
User1	0.96
User5	0.92
User4	0.80

表 13: 余弦相似度

用户	平均打分
User3	$\frac{(1+3+1+1)}{4} = 1.5$
User1	$\frac{(1+1+5+3)}{4} = 2.5$
User5	$\frac{(2+2+2+4)}{4} = 2.5$
User2	$\frac{(3+5+4)}{3} = 4$

表 14: 平均打分

source, pasta, coffee,milk,water,juices放在一起销售来提高销售量。

按照支持度排序的结果如图3所示，其中前5条规则为：

$$\begin{aligned} &\{milk\} \\ &\{pasta\} \rightarrow \{milk\} \\ &\{water\} \rightarrow \{milk\} \\ &\{biscuits\} \rightarrow \{milk\} \\ &\{biroches\} \rightarrow \{milk\} \end{aligned} \tag{3}$$

从图中可以看到milk的支持度为100，$\{pasta\} \rightarrow \{milk\}$规则为冗余的因为其提升度是小于1的二者负相关，二者单独放在一起并不能共同提高销量。

按照置信度排序的结果如图4所示，其中前5条规则为：

$$\begin{aligned} &\{biscuits, pasta\} \rightarrow \{milk\} \\ &\{pasta, water\} \rightarrow \{milk\} \\ &\{juices\} \rightarrow \{milk\} \\ &\{tomatosource\} \rightarrow \{pasta\} \\ &\{yoghurt\} \rightarrow \{milk\} \end{aligned} \tag{4}$$

因为这5条规则的提升度都大于1，因此可以将tomato source, pasta, coffee,milk,water,yoghurt放在一起销售来提高销售量。

Sort by: Lift

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2
5363	11.600	53.200	1.520	pasta	tomato sou...	
3466	7.500	45.500	1.300	pasta	coffee	milk
3590	7.800	57.900	1.254	milk	biscuits	pasta
4417	9.600	56.000	1.214	milk	water	pasta
3818	8.300	53.300	1.155	milk	juices	
7045	15.200	52.200	1.131	milk	yoghurt	
9468	20.500	51.500	1.117	milk	biscuits	
5363	11.600	51.300	1.112	milk	tomato sou...	
6949	15.000	49.900	1.081	milk	coffee	
7084	15.300	49.700	1.077	milk	brioches	
4951	10.700	47.300	1.025	milk	coke	
12879	27.900	46.700	1.012	milk	water	
4803	10.400	46.400	1.006	milk	tunny	
46243	100.000	46.100	1.000	milk		
16201	35.000	45.900	0.994	milk	pasta	
5050	10.900	45.400	0.985	milk	beer	

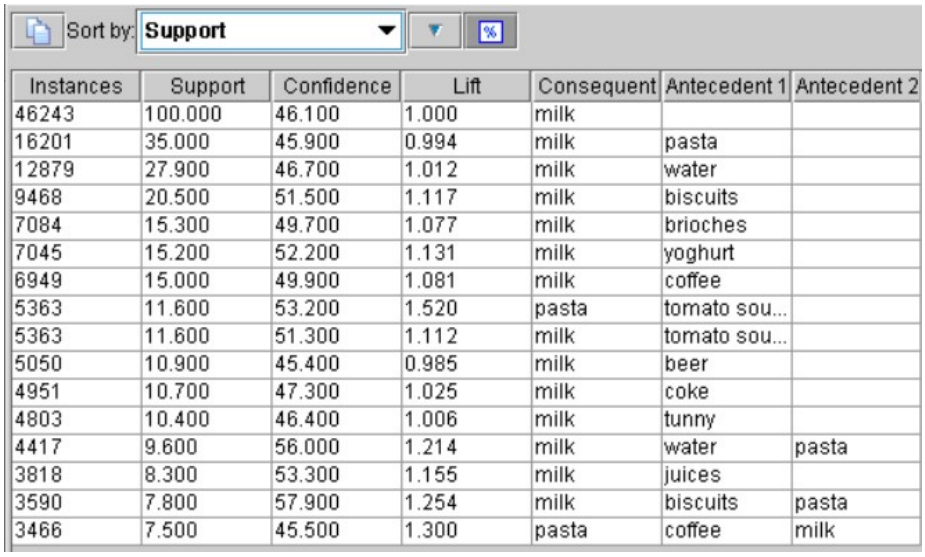
图 2: sort by lift

1.5.2 Question2

1. 如图(a),(b),(c)三种情况分别如图5，图6，图7所示。

2. 选择第二种Minimum records per child branch=15，因为从混淆矩阵中可以看出这个参数的决策树正确率最高。

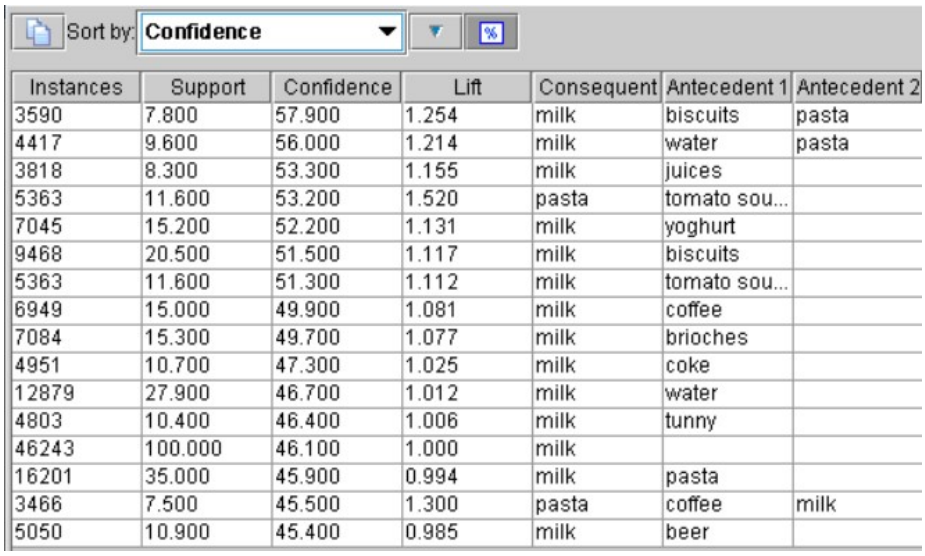
3. 使用Minimum records per child branch=15的决策树模型，得到的预测结果如表15所示。



Sort by: **Support**

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2
46243	100.000	46.100	1.000	milk		
16201	35.000	45.900	0.994	milk	pasta	
12879	27.900	46.700	1.012	milk	water	
9468	20.500	51.500	1.117	milk	biscuits	
7084	15.300	49.700	1.077	milk	brioches	
7045	15.200	52.200	1.131	milk	yoghurt	
6949	15.000	49.900	1.081	milk	coffee	
5363	11.600	53.200	1.520	pasta	tomato sou...	
5363	11.600	51.300	1.112	milk	tomato sou...	
5050	10.900	45.400	0.985	milk	beer	
4951	10.700	47.300	1.025	milk	coke	
4803	10.400	46.400	1.006	milk	tunny	
4417	9.600	56.000	1.214	milk	water	pasta
3818	8.300	53.300	1.155	milk	juices	
3590	7.800	57.900	1.254	milk	biscuits	pasta
3466	7.500	45.500	1.300	pasta	coffee	milk

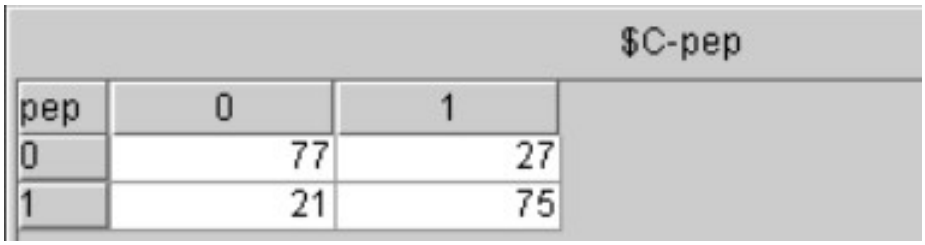
图 3: sort by support



Sort by: **Confidence**

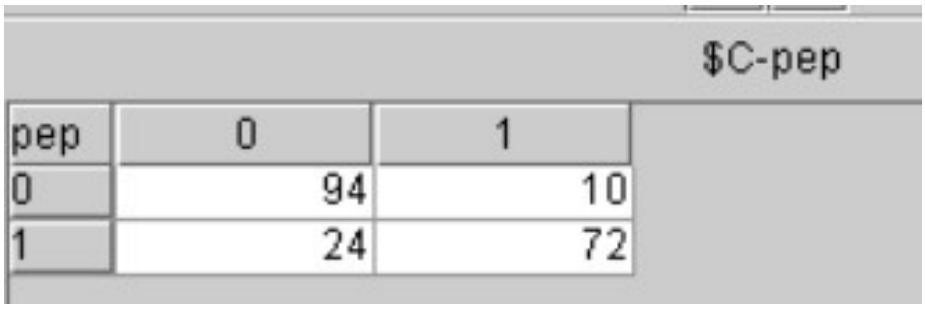
Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2
3590	7.800	57.900	1.254	milk	biscuits	pasta
4417	9.600	56.000	1.214	milk	water	pasta
3818	8.300	53.300	1.155	milk	juices	
5363	11.600	53.200	1.520	pasta	tomato sou...	
7045	15.200	52.200	1.131	milk	yoghurt	
9468	20.500	51.500	1.117	milk	biscuits	
5363	11.600	51.300	1.112	milk	tomato sou...	
6949	15.000	49.900	1.081	milk	coffee	
7084	15.300	49.700	1.077	milk	brioches	
4951	10.700	47.300	1.025	milk	coke	
12879	27.900	46.700	1.012	milk	water	
4803	10.400	46.400	1.006	milk	tunny	
46243	100.000	46.100	1.000	milk		
16201	35.000	45.900	0.994	milk	pasta	
3466	7.500	45.500	1.300	pasta	coffee	milk
5050	10.900	45.400	0.985	milk	beer	

图 4: sort by sconfidence



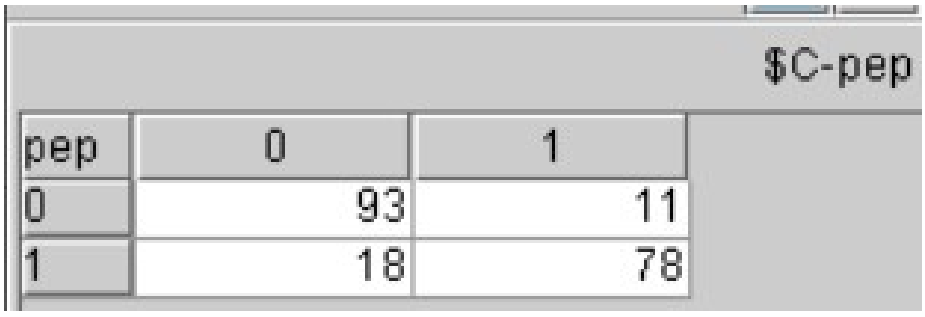
	$C-pep	
pep	0	1
0	77	27
1	21	75

图 5: Minimum records per child branch=56



	$C-pep	
pep	0	1
0	94	10
1	24	72

图 6: Minimum records per child branch=15



	$C-pep	
pep	0	1
0	93	11
1	18	78

图 7: Minimum records per child branch=10

age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
22	0	1	14000	0	3	0	1	1	0	0
47	0	0	16700	1	1	0	1	1	0	1
54	1	1	43400	1	1	1	1	1	0	1
65	1	2	60000	1	1	0	1	1	0	1
37	0	0	27700	0	1	1	0	0	0	1
44	0	0	38784	1	0	0	1	1	0	0
20	1	0	10200	1	0	0	1	1	1	0
46	0	0	22000	1	1	1	1	0	1	1
40	1	1	37400	1	2	0	1	1	0	1

表 15: Appendix1