

 -CAL-FENGHAI.Network Information Technology Company.2020YEAR

**数据挖掘期末考试在线测试答案**

一个食品连锁店每周的事务记录如下表所示，其中每一条事务表示在一项收款机业务中卖出的项目，假定supmin=20%，confmin=40%，使用Apriori算法计算生成的关联规则，标明每趟数据库扫描时的候选集和大项目集。

|  |  |  |  |
| --- | --- | --- | --- |
| **事务** | **项目** | **事务** | **项目** |
|     T1    T2                       T3 | 面包、果冻、花生酱面包、花生酱面包、牛奶、花生酱 |     T4    T5 | 啤酒、面包啤酒、牛奶 |

解：1) 扫描数据库对每个候选计算支持

|  |
| --- |
| C1 |
| 项集 | 支持度 |
| {面包}{花生酱}{牛奶}{啤酒}{果冻} | 4/53/52/52/51/5 |

2) 比较候选支持度与最小支持度，得出频繁项集L1

|  |
| --- |
| L1 |
| 项集 | 支持度 |
| {面包}{花生酱}{牛奶}{啤酒}{果冻} | 4/53/52/52/51/5 |

3)由L1产生候选C2

|  |
| --- |
| C2 |
| 项集 |
| {面包，花生酱}{面包，牛奶}{面包，啤酒}{面包，果冻}{花生酱，牛奶}{花生酱，啤酒}{花生酱，果冻}{牛奶，啤酒}{牛奶，果冻}{啤酒，果冻} |

4）扫描，对每个候选计算支持度

|  |
| --- |
| C2 |
| 项集 | 支持度 |
| {面包，花生酱}{面包，牛奶}{面包，啤酒}{面包，果冻}{花生酱，牛奶}{花生酱，啤酒}{花生酱，果冻}{牛奶，啤酒}{牛奶，果冻}{啤酒，果冻} | 3/51/51/51/51/501/51/500 |

5)比较候选支持度与最小支持度，得出频繁项集L2

|  |
| --- |
| L2 |
| 项集 | 支持度 |
| {面包，花生酱}{面包，牛奶}{面包，啤酒}{面包，果冻}{花生酱，牛奶}{花生酱，果冻}{牛奶，啤酒} | 3/51/51/51/51/51/51/5 |

6)由L2产生候选C3

|  |
| --- |
| C3 |
| 项集 |
| {面包，花生酱，牛奶}{面包，花生酱，啤酒}{面包，花生酱，果冻}{面包，牛奶，啤酒}{面包，牛奶，果冻}{面包，啤酒，果冻}{花生酱，牛奶，果冻}{花生酱，牛奶，啤酒} |

7）扫描，对每个候选计算支持度

|  |
| --- |
| C3 |
| 项集 | 支持度 |
| {面包，花生酱，牛奶}{面包，花生酱，啤酒}{面包，花生酱，果冻}{面包，牛奶，啤酒}{面包，牛奶，果冻}{面包，啤酒，果冻}{花生酱，牛奶，果冻}{花生酱，牛奶，啤酒} | 1/501/500000 |

8）比较候选支持度与最小支持度，得出频繁项集L3

|  |
| --- |
| C3 |
| 项集 | 支持度 |
| {面包，花生酱，牛奶}{面包，花生酱，果冻} | 1/51/5 |

下面计算关联规则：

<1>{面包，花生酱，牛奶}的非空子集有{面包，花生酱}，{面包，牛奶}，{花生酱，牛奶}，{面包}，{花生酱}，{牛奶}

{面包，花生酱} {牛奶} confidence==33.3%

{面包，牛奶} {花生酱} confidence==100%

{花生酱，牛奶} {面包} confidence==100%

{面包} {花生酱，牛奶} confidence==25%

{花生酱} {面包，牛奶} confidence==33.3%

{牛奶} {面包，花生酱} confidence==50%

故强关联规则有{面包，牛奶} {花生酱}，{花生酱，牛奶} {面包}，

{牛奶} {面包，花生酱}

<2>{面包，花生酱，果冻}的非空子集有{面包，花生酱}，{面包，果冻}，{花生酱，果冻}，{面包}，{花生酱}，{果冻}

{面包，花生酱} {果冻} confidence==33.3%

{面包，果冻} {花生酱} confidence==100%

{花生酱，果冻} {面包} confidence==100%

{面包} {花生酱，果冻 } confidence==25%

{花生酱} {面包，果冻} confidence==33.3%

{果冻} {面包，花生酱} confidence=100%

故强关联规则有{面包，果冻} {花生酱}，{花生酱，果冻} {面包}，

{果冻} {面包，花生酱}

The following shows a history of customers with their incomes, ages and an attribute called “Have\_iPhone”indicating whether they have an iPhone. We also indicate whether they will buy an iPad or not in the lastcolumn.

No.   Income      Age     Have\_iPhone   Buy\_iPad1        high        young       yes               yes2        high        old           yes               yes3        medium   young       no                yes4        high        old            no                yes5        medium   young       no                no6        medium   young       no                no7        medium   old           no                no8        medium   old           no                no(a) We want to train a CART decision tree classifier to predict whether a new customer will buy an iPad or not. We define the value of attribute Buy\_iPad is the label of a record.(i) Please find a CART decision tree according to the above example. In the decision tree, wheneverwe process a node containing at most 3 records, we stop to process this node for splitting.(ii) Consider a new young customer whose income is medium and he has an iPhone. Please predictwhether this new customer will buy an iPad or not.(b) What is the difference between the C4.5 decision tree and the ID3 decision tree Why is there a difference?

解：解：a.(i)对于所给定样本的期望信息是：-log2-log2=1

属性Income的样本:

Info(high)=-3 log21-0 log20=0

Info(medium)=-  log2- log2=0.72193

期望信息为：×0+×0.72193=0.27072

信息增益为：Gain（Income）=1-E(Income)= 0.729277

同样计算知：

Gain(Age)=0.09436

Gain(Have\_iPhone)=0.311

 这三个属性中Income的Gain最大，所以选择Income为最优特征，于是根节点生成两个子节点，一个是叶节点，对另一个节点继续使用以上方法，在A2，A3选择最优特征及其最优切分点，结果是Age。依此计算得，CART树为：

Young

Old

medium

High

（ii）这个新的年轻、中等收入、有IPhone的顾客，将不会购买IPad。

（b）C4.5决策树算法和ID3算法相似，但是C4.5决策树算法是对ID3算法的改进，ID3算法在生成决策树的过程中，使用信息增益来进行特征选择，是选择信息增益最大的特征；C4.5算法在生成决策树的过程中，用信息增益比来选择特征，是选择信息增益比最大的特征。因为信息增益的大小是相对于训练数据集而言的，并没有绝对的意义，在分类困难时，也就是在训练数据集的经验熵大的时候，信息增益会偏大，反之，信息增益会偏小。使用信息增益比可以对这一问题进行校正。

Consider the following eight two-dimensional data points:x1: (23, 12), x2: (6, 6), x3: (15, 0), x4: (15, 28), x5:(20, 9), x6: (8, 9), x7: (20, 11), x8: (8, 13),Consider algorithm k-means.Please answer the following questions. You are required to show the information about each final cluster(including the mean of the cluster and all data points in this cluster). You can consider writing a program forthis part but you are not required to submit the program.(a) If k = 2 and the initial means are (20, 9) and (8, 9), what is the output of the algorithm?(b) If k = 2 and the initial means are (15, 0) and (15, 29), what is the output of the algorithm?

解：(a)已知K=2，初始质心是(20, 9)、(8, 9)

 则：

|  |  |  |  |
| --- | --- | --- | --- |
| M1 | M2 | K1 | K2 |
| (20, 9) | (8, 9) | (20,9),(23,12),(15,0), (15,28), (20,11) |  (8,9), (6,6), (8,13) |
| (18.6,12) | (7.3,9.3) | (23,12),(15,28),(20,9),(20,11)} | (15,0),(6,6),(8,9),(8,13) |
| (19.5,15) | (9.5,7) | (23,12),(15,28),(20,9),(20,11) | (15,0),(6,6),(8,9),(8,13) |

所以，算法输出两个簇：

 K1={x1,x4,x5,x7}

 K2={x2,x3,x6,x8}

（b）已知K=2，初始质心是(15, 0)、(15, 29)

则：

|  |  |  |  |
| --- | --- | --- | --- |
| M1 | M2 | K1 | K2 |
| (15, 0) | (15, 29) | (23,12),(6,6),(15,0),(20,9),(8,9),(20,11),(8,13) | (15,28) |
| (14.3,8.6) | (15,28) | (23,12),(6,6),(15,0),(20,9),(8,9),(20,11),(8,13) | (15,28) |

所以，算法输出两个簇：

 K1={x1,x2,x3,x5,x6,x7,x8}

K2={x4}

4. Consider eight data points

The following matrix shows the pairwise distances between any two points.

 1 2 3 4 5 6 7 8

1 0

2 11 0

3 5 13 0

4 12 2 14 0

5 7 17 1 18 0

6 13 4 15 5 20 0

7 9 15 12 16 15 19 0

8 11 20 12 21 17 22 30 0

Please use the agglomeration approach to cluster these eight points into two groups/clusters by using distance complete linkage.

Please write down all data points for each cluster and write down the distance between the two clusters.

3 5距离1合并为簇（3，5）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 　 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0 | 　 | 　 | 　 | 　 | 　 | 　 |  |
| 2 | 11 | 0 | 　 | 　 | 　 | 　 | 　 |  |
| 3 | 5 | 13 | 0 | 　 | 　 | 　 | 　 |  |
| 4 | 12 | 2 | 14 | 0 | 　 | 　 | 　 |  |
| 5 | 7 | 17 | 1 | 18 | 0 |  |  |  |
| 6 | 13 | 4 | 15 | 5 | 20 | 　 0  | 　 |  |
| 7 | 9 | 15 | 12 | 16 | 15 | 19 | 　 0 |  |
| 8 | 11 | 20 | 12 | 21 | 17 | 22 | 30 | 0 |

2 4距离2合并为簇（2，4）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3,5 | 4 | 6 | 7 | 8 |
| 1 | 0 |  |  |  |  |  |  |
| 2 | 11 | 0 |  |  |  |  |  |
| 3,5 | 5 | 13 | 0 |  |  |  |  |
| 4 | 12 | 2 | 14 | 0 |  |  |  |
| 6 | 13 | 4 | 15 | 5 | 0 |  |  |
| 7 | 9 | 15 | 12 | 16 | 19 | 0 |  |
| 8 | 11 | 20 | 12 | 21 | 22 | 30 | 0 |

（2 ，4）6距离4合并为簇（2，4,6）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 　 | 1 | 2,4 | 3,5 | 6 | 7 | 8 |
| 1 | 0 | 　 | 　 | 　 | 　 |  |
| 2,4 | 11 | 0 | 　 | 　 | 　 |  |
| 3,5 | 5 | 13 | 0 | 　 | 　 |  |
| 6 | 13 | 4 | 15 | 0 |  |  |
| 7 | 9 | 15 | 12 | 19 | 　 0 |  |
| 8 | 11 | 20 | 12 | 22 | 30 | 0 |

1距离（3,5）为5合并为簇（1，3,5）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 　 | 1 | 2,4,6 | 3,5 | 7 | 8 |
| 1 | 0 | 　 | 　 | 　 |  |
| 2,4,6 | 11 | 0 | 　 | 　 |  |
| 3，5 | 5 | 13 | 0 |  |  |
| 7 | 9 | 15 | 12 | 　 0 |  |
| 8 | 11 | 20 | 12 | 30 | 0 |

（1,3,5）距离7为9合并为簇（1，3,5,7）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | 1，3,5 | 2,4,6 | 7 | 8 |
| 1,3,5 | 0 | 　 | 　 |  |
| 2，4,6 | 11 | 0 | 　 |  |
| 7 | 9 | 15 | 0 |  |
| 8 | 11 | 20 | 30 | 0 |

（1,3,5,7） 距离8为11合并为簇（1,3,5,7，8）

|  |  |  |  |
| --- | --- | --- | --- |
| 　 | 1，3,5,7 | 2,4,6 | 8 |
| 1，3,5,7 | 0 | 　 |  |
| 2,4,6 | 11 | 0 |  |
| 8 | 30 | 20 | 0 |

0 合并

|  |  |  |
| --- | --- | --- |
|  | 1，3，5,7，8 | 2,4,6 |
| 1,3,5,7，8 | 0 |  |
| 2,4,6 | 11 | 0 |