数据仓库与数据挖掘复习题

1.假设数据挖掘的任务是将如下的8个点(用(x,y)代表位置)聚类为 3个类：

X1(2,10)、X2(2,5)、X3(8,4)、X4(5,8)、X5(7,5)、X6(6,4)、X7(1,2)、X8(4,9),距离选 择欧几里德距离.假设初始选择X1(2,10)、X4(5,8)、X7(1,2)为每个聚类的中央,请

用K\_means算法来计算：

(1)在第一次循环执行后的3个聚类中央；

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 答：第一 | 一次迭代： | 中央点1: | X1(2, 10), | 2: X4(5, 8), X7(1, | | 2) | | |
|  | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 1 | 0 | 25 | 36+36 | 9+4 | 25+25 | 16+36 | 1+64 | 4+1 |
| 2 | 9+4 | 9+9 | 9+16 | 0 | 4+9 | 1+16 | 16+36 | 1+1 |
| 3 | 1+64 | 1+9 | 53 | 16+36 | 45 | 29 | 0 | 58 |

答案：在第一次循环执行后的3个聚类中央：

1. X1(2, 10)
2. X3, X4, X5, X6, X8 (6, 6)
3. X2, X7 (1.5, 3.5)

(2)经过两次循环后,最后的3个族分别是什么?

第二次迭代：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| d2 | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 1 | 0 | 25 | 36+36 | 9+4 | 25+25 | 16+36 | 1+64 | 4+1 |
| 2 | 32 | 17 | 8 | 5 | 2 | 4 | 41 | 1+1 |
| 3 | 52+6.52 | 52+1.52 | 6.52+0.52 | 3.52+4.52 | 5.52+1.52 | 4.52+0.52 | 0.52+1.52 | 2.52+5.52 |

答案：1: X1, X8 (3.5, 9.5)

1. X3, X4, X5, X6 (6.5, 5.25)
2. X2, X7(1.5, 3.5)

2.数据库有 4个事务.设 min\_sup=60%,min\_conf=80%.

|  |  |  |
| --- | --- | --- |
| TID | data | Transaction |
| T100 | 6/6/2007 | K,A,D,B |
| T200 | 6/6/2007 | D,A,C,E,B |
| T300 | 6/7/2007 | C,A,B,E |
| T400 | 6/10/2007 | B,A,D |

a使用Apriori算法找出频繁项集,并写出具体过程

(a)Apriori 算法:

~~{K}1~~{A}4{A,B}4{A,B,D) 3

{A}4{B}4{A,D}3

{B}4{D}3{B,D}3

{D}3

{€}—2-

{E}—2-

频繁项集为3项集{A,B,D}:3

b.列出所有的强关联规那么,使它们与下面的元规那么匹配,其中, X是代表顾客

的变量,itemi是表示项的变量(例如,“A〞、“B〞等)：

Vx=transac,tior( b,uys)X item *,* buy)sX item, 3*b)* i[sy(s| X item

答：所有频繁子项集有{A},{B},{D},{A,B},{A,D},{B,D}

AAB=>D conf=3/4=75%

AAD=>B conf=3/3=100% V

BAD=>A conf=3/3=100%V

因此,满足条件的强关联规那么有：

AAD=>B{supp=75%,conf=100%}

BAD=>A{supp=75%,conf=100%}

1.给定如下的数据库表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Sky | AirTemp | Humidity | Wind | Water | Forecast | Enjoysport |
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Warm | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | yes |

请计算属性Sky的信息增益.

答：

C1 : Enjoysport=yes=3

C2 : Enjoysport=no=1

I(yes,no)=-3/4log 2 3/4-1/4log 2 1/4=0.811

|  |  |  |
| --- | --- | --- |
| sky | C1 | C2 |
| rainy | 0 | 1 |
| sunny | 3 | 0 |

I(sky)=1/4I(0,1)+3/4I(3,0)=0

Gain(sky)=0.811

习题：

1.以汽车保险为例：假定练习数据库具有两个属性：年龄和汽车类型.

年龄序数属性

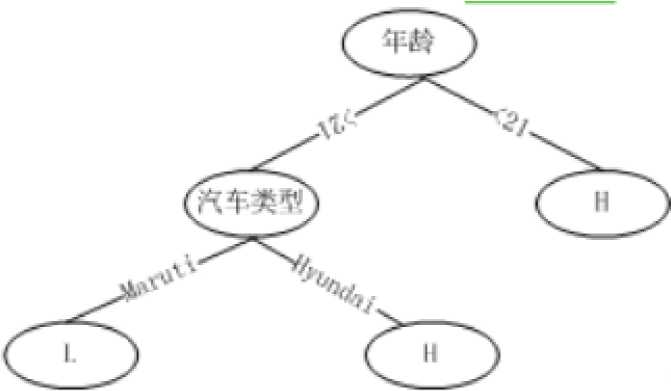
汽车类型一一分类属性

-L:低〔风险〕,H:高〔风险〕

|  |  |  |
| --- | --- | --- |
| 年龄 | 汽车类型 | 类 |

|  |  |  |
| --- | --- | --- |
| >21 | Maruti | L |
| >21 | Hyundai | H |
| <21 | Maruti | H |
| <21 | Indica | H |
| >21 | Maruti | L |
| >21 | Hyundai | H |

使用ID3算法得到一个决策树.



2.下面是一个超市某商品连续24个月的销售数据〔单位：百万元〕：

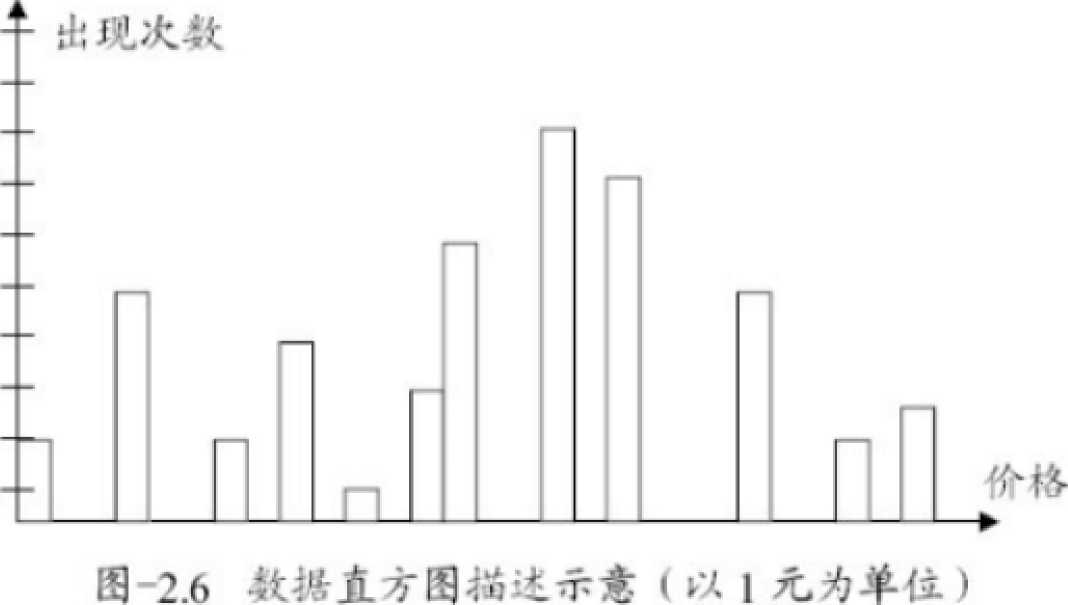
21,16,21, 19, 24, 27, 23, 22, 21, 20, 17, 16, 20, 23, 22, 18, 24, 26, 25,

20, 26, 23, 21, 15, 17

请使用等深、等宽和自定义区间的方法对数据进行分箱,做出利

用各种分箱方法得到的直方图.

上述数据所涉成属性彳酊算率时的直方图如图-2.6所示.



3.数据库有4个事务.设 min\_sup = 60%, min\_conf = 80%.使用

Apriori算法找出所有的频繁项集,并针对每个频繁项集构造强关 联规那么,列出每个规那么的支持度和置信度.

|  |  |
| --- | --- |
| hd | Dateiiems\_bought |
| T100  T200  T300  T400 | 10/15,W{KA.D.B)  10.15/99{D,A,C,E,B)  101999{CAB-E)  *1W2y99*(BAD} |

(b)Apriori 算法：

{A,B} 4{A,B,D} 3

{A,D} 3

{B,D} 3

{KF^{A}4

{A} 4{B}4

{B} 4{D}3

{D} 3

{€}—2- {E}—2-

频繁项集为3项集{A,B,D}:3

所有频繁子项集有{A},{B},{D},{A,B},{A,D},{B,D}

AAB=>D conf=3/4=75%*乂*

AAD=>B conf=3/3=100%V

BAD=>A conf=3/3=100%V

因此,满足条件的强关联规那么有：

AAD=>B{supp=75%,conf=100%}

BAD=>A{supp=75%,conf=100%}