[商务智能]2016真题

一、 问答（4*10）

1. 举例说明支持度与置信度的概念及计算方式 PPTch3关联分析P7-10

支持度：指项集X在记录集合D中出现的概率

Support(支持度): probability that a transaction contains X

sup(X)=|X|/n

置信度：项集X出现的情况下，项集Y在记录集合D中同时出现的条件概率，即X并Y的支持度除以X的支持度

confidence (置信度)： conditional probability that a transaction having X also contains Y

Conf(X ® Y)=|XY| / |X|=sup(XY) / sup(X)

2. 说明分类与聚类的不同之处

Classification (分类)，对于一个classifier，通常需要你告诉它“这个东西被分为某某类”这样一些例子，理想情况下，一个 classifier 会从它得到的训练集中进行“学习”，从而具备对未知数据进行分类的能力，这种提供训练数据的过程通常叫做supervised learning (监督学习)，

Clustering (聚类)，简单地说就是把相似的东西分到一组，聚类的时候，我们并不关心某一类是什么，我们需要实现的目标只是把相似的东西聚到一起。因此，一个聚类算法通常只需要知道如何计算相似度就可以开始工作了，因此 clustering 通常并不需要使用训练数据进行学习，这在Machine Learning中被称作unsupervised learning (无监督学习).

3. 说明数据仓库与数据库的不同之处

数据仓库是一个面向主题的，集来成的，随时间变化，稳定的用于支持组织决策的数据集合。

数据集市是一种部门级的数据仓库，它包含的数据量较少，是面向一个部门的分析需求而建立的。

主要区别：

数据来源不同：数据仓库的数据来源于遗留系统、OLTP系统、外部数据，数据集市的数据来源于数据仓库。

范围不同：数据仓库为企业级的数据仓库，数据集市是部门级或工作组级的数据仓库。

主题不同：数据仓库的主题为企业主题，数据集市的主题为部门或特殊的分析主题。

数据粒度不同：数据仓库的数据粒度最细，数据集市的数据粒度较粗。

数据结构不同：数据仓库的数据结构为规范化结构（第3范式），数据集市的数据结构为星型模式、雪片模式、事实星座。

历史数据需求量不同：数据仓库比数据集市需要更多的历史数据。

优化方向不同：数据仓库要便于处理海量数据和数据探索，数据集市要便于访问和分析快速查询。

4. 举例说明多维数据分析的主要操作类型有哪些

切片、 切块、向上钻取、向下钻取、钻透、旋转

列如一个产品，日期，国家的一个三维数据表

切片：固定国家维度取一个固定值，其他维不变得到的立方体称为一个切片

切块：固定国家维度取多个值，其他维不变得到的立方体称为一个切块

上钻：（国家，产品，季度）向上钻取汇总（国家，产品，年）或者（国家，产品）

下钻：（国家，产品，年）向下展示下一层的更细的数据（国家，产品，季度）

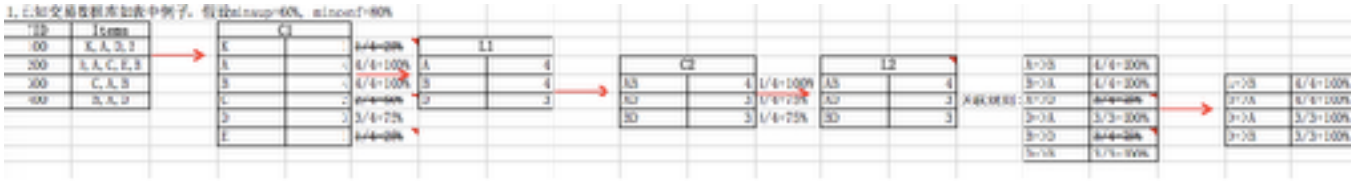
旋转：行列互换提供不同的视角(product, year) 换为(location, year)

二、 计算（2*30）

1. 已知交易数据库如表中例子，假设minsup=60%，minconf=80%

TID	Items
100	K, A, D, B
200	D, A, C, E, B
300	C, A, B
400	B, A, D

请用Apriori算法找出所有的频繁项集以及关联规则　17高婧，17易龙飞



频繁项集：

C1:

{A}=100%

{B}=100%

{C}=2/4

{D}=3/4

{E}=1/4

{K}=1/4

L1:

{A},{B},{D}

C2:

{AB}=100%

{AD}=3/4

{BD}=3/4

L2:

{AB},{AD},{BD}

C3:

{ABD}=3/4

L3:

{ABD}

关联规则：

>B=sup(AB)/sup(A)=1/1=1>80%

>D=sup(AD)/sup(A)=3/4<80%

>D=3/4<80%

B->A=1>80%

D->A=sup(AD)/sup(D)=1>80%

D->B=sup(BD)/sup(D)=1>80%
AB->D=sup(ABD)/sup(AB)=3/4<80%
AD->B=sup(ABD)/sup(AD)=1>80%
BD->A=sup(ABD)/sup(BD)=1>80%
综上：关联规则是A->B,B->A,D->A,D->B,AD->B,BD->A

2. 如下图所示有关温度及压力的测试记录情况
TID 温度 压力 警告1 警告2 警告3

TID	温度	压力	警告1	警告2	警告3
1	95	1105	0	0	1
2	85	1040	1	1	0
3	103	1090	1	1	1
4	97	1084	1	0	0
5	80	1038	0	1	1
6	100	1080	1	1	0
7	83	1025	1	0	1
8	86	1030	1	0	0
9	101	1100	1	1	1

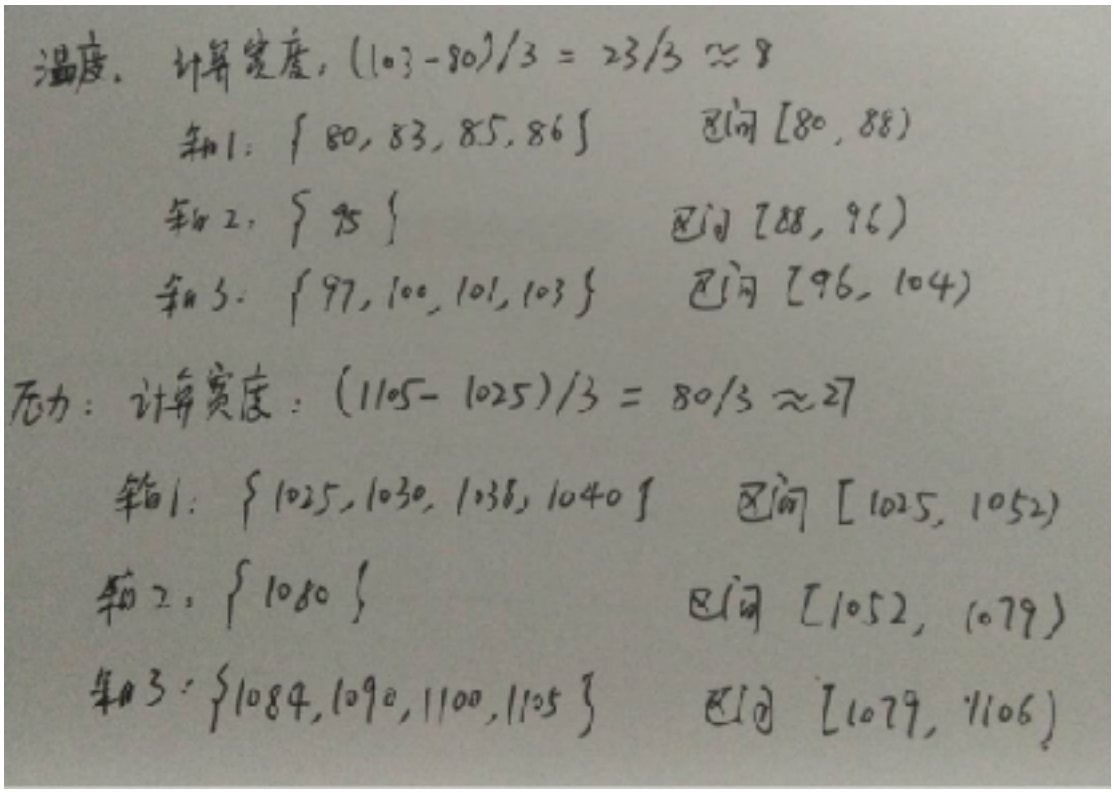
假设对该数据集的连续属性采用如下离散化方法：
（1） 将每个连续属性的值域划分为三个等宽的箱

（2） 将每个连续属性的值域划分为三个箱，每个箱子包含的事务个数相同
针对两个不同的离散化方法，分别构造数据集的二元化属性值(0, 1)的数据集合

解：（1）
（2）将每个连续属性的值域划分为三个箱，每个箱子包含的事务个数相同
针对两个不同的离散化方法，分别构造数据集的二元化属性值(0, 1)的数据集合
解：因为需确保连续的属性值划分三个箱子，每个箱子事务个数相同，故采用等频率分箱。

TID	警告1	按警告1取值划分	TID	警告1
1	0		1	0
2	1		5	0
3	1		2	1
4	1		3	1
5	0		4	1

6	1		6	1
7	1		7	1
8	1		8	1
9	1		9	1



eg:温度	整数值	x1	x2
箱1.[80,83,85]	0	0	0
箱2.[86,95,97]	1	0	1
箱3.[100,101.103]	2	1	0

警告1、[0,0,1], [1,1,1], [1,1,1]

TID	警告2	按警告2取值划分	TID	警告2
1	0		1	0
2	1		4	0
3	1		7	0
4	0		8	0
5	1		2	1

6	1		3	1
7	0		5	1
8	0		6	1
9	1		9	1

警告2、[0,0,0],[0,1,1],[1,1,1]

TID	警告3	按警告3取值划分	TID	警告3
1	1		2	0
2	0		4	0
3	1		6	0
4	0		8	0
5	1		1	1
6	0		3	1
7	1		5	1
8	0		7	1
9	1		9	1

警告3、[0,0,0],[0,1,1],[1,1,1]

知识点：书Page82
离散化方法分:有监督(分箱法)和无监督(基于熵方法【自顶向下】、基于卡方统计方法ChiMerge【自底向上】)。
分箱离散化分为等距离（又称宽度分箱）和等频率（又称深度分箱）。
宽度分箱：将每个取值映射到等大小的区间方法。（若区间个数为k,每个区间=给定属性的（最大值-最小值）/k)。
深度分箱：将每个取值映射到一个区间，每个区间包含的取值个数大致相同。

二元化

一种分类属性二元化的简单技术如下：如果有m个分类值，则将每个原始值唯一地赋予区间[0, m 1]中的一个整数。如果属性是有序的，则赋值必须保持序关系。（注意，即使属性原来就

用整数表示，但如果这些整数不在区间[0, m 1]中，则该过程也是必需的。）然后，将这m个整数的每一个都变换成一个二进制数。由于需要n = log2m 个二进位表示这些整数，因此要使用n个二元属性表示这些二进制数。例如，一个具有5个值{awful, poor, OK, good, great}的分类变量需要三个二元变量x1、x2、x3。转换见表2-5。

表2-5 一个分类属性到三个二元属性的变换

分类值	整数值	x1	x2	x3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

这样的变换可能导致复杂化，如无意之中建立了转换后的属性之间的联系。例如，在表2-5中，属性x2和x3是相关的，因为good值使用这两个属性表示。此外，关联分析需要非对称的二元属性，其中只有属性的出现（值为1）才是重要的。因此，对于关联问题，需要为每一个分类值引入一个二元属性，如表2-6所示。如果结果属性的个数太多，则可以在二元化之前使用下面介绍的技术减少分类值的个数。

表2-6 一个分类属性到五个非对称二元属性的转换

分类值	整数值	x1	x2	x3	x4	x5

awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

同样，对于关联问题，可能需要用两个非对称的二元属性替换单个二元属性。考虑记录人的性别（男、女）的二元属性，对于传统的关联规则算法，该信息需要转换成两个非对称的二元属性，其中一个仅当是男性时为1，而另一个仅当是女性时为1。（对于非对称的二元属性，由于其提供一个二进制位信息需要占用存储器的两个二进制位，因而在信息的表示上不太有效。