一、单选题（共80题）

( D )的目的缩小数据的取值范围，使其更适合于数据挖掘算法的需要，并且能够得到和原始数据相同的分析结果。

A.数据清洗       B.数据集成

C.数据变换       D.数据归约

某超市研究销售纪录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？(A)  
   A. 关联规则发现       B. 聚类  
   C. 分类               D. 自然语言处理

以下两种描述分别对应哪两种对分类算法的评价标准？(A)  
  (a)警察抓小偷，描述警察抓的人中有多少个是小偷的标准。  
  (b)描述有多少比例的小偷给警察抓了的标准。  
  A. Precision,Recall   B.Recall,Precision  
  A. Precision,ROC      D. Recall,ROC

将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务？(C)  
  A. 频繁模式挖掘       B. 分类和预测

C. 数据预处理         D. 数据流挖掘

当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？(B)  
  A. 分类               B. 聚类

C. 关联分析           D. 隐马尔可夫链

建立一个模型，通过这个模型根据已知的变量值来预测其他某个变量值属于数据挖掘的哪一类任务？(C)  
  A. 根据内容检索       B. 建模描述  
  C. 预测建模           D. 寻找模式和规则

下面哪种不属于数据预处理的方法？ (D)  
  A.变量代换            B.离散化

C.聚集                D.估计遗漏值

假设12个销售价格记录组已经排序如下：5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 使用如下每种方法将它们划分成四个箱。等频（等深）划分时，15在第几个箱子内？ (B)  
  A.第一个             B.第二个

C.第三个             D.第四个

下面哪个不属于数据的属性类型：(D)  
  A.标称               B.序数

C.区间               D.相异

只有非零值才重要的二元属性被称作：( C )  
  A.计数属性           B.离散属性

C.非对称的二元属性   D.对称属性

以下哪种方法不属于特征选择的标准方法：    (D)  
  A.嵌入               B.过滤

C.包装               D.抽样

下面不属于创建新属性的相关方法的是： (B)  
 A.特征提取            B.特征修改

C.映射数据到新的空间  D.特征构造

下面哪个属于映射数据到新的空间的方法？ (A)  
 A.傅立叶变换          B.特征加权

C.渐进抽样            D.维归约

假设属性income的最大最小值分别是12000元和98000元。利用最大最小规范化的方法将属性的值映射到0至1的范围内。对属性income的73600元将被转化为：(D)  
 A.0.821              B.1.224

C.1.458              D.0.716

一所大学内的各年纪人数分别为：一年级200人，二年级160人，三年级130人，四年级110人。则年级属性的众数是： (A)  
 A.一年级              B.二年级

C.三年级              D.四年级

下列哪个不是专门用于可视化时间空间数据的技术：(B)  
 A.等高线图            B.饼图

C.曲面图              D.矢量场图

在抽样方法中，当合适的样本容量很难确定时，可以使用的抽样方法是： (D)  
 A.有放回的简单随机抽样

B.无放回的简单随机抽样

C.分层抽样

D 渐进抽样

数据仓库是随着时间变化的,下面的描述不正确的是 (C)  
A.数据仓库随时间的变化不断增加新的数据内容  
B.捕捉到的新数据会覆盖原来的快照  
C.数据仓库随事件变化不断删去旧的数据内容  
D.数据仓库中包含大量的综合数据,这些综合数据会随着时间的变化不断地进行重新综合

下面关于数据粒度的描述不正确的是: (C)  
A.粒度是指数据仓库小数据单元的详细程度和级别  
B.数据越详细,粒度就越小,级别也就越高  
C.数据综合度越高,粒度也就越大,级别也就越高  
D.粒度的具体划分将直接影响数据仓库中的数据量以及查询质量

有关数据仓库的开发特点,不正确的描述是: (A)  
A.数据仓库开发要从数据出发  
B.数据仓库使用的需求在开发出去就要明确  
C.数据仓库的开发是一个不断循环的过程,是启发式的开发  
D.在数据仓库环境中,并不存在操作型环境中所固定的和较确切的处理流,数据仓库中数据分析和处理更灵活,且没有固定的模式

关于OLAP的特性,下面正确的是: (D)  
(1)快速性 (2)可分析性 (3)多维性 (4)信息性 (5)共享性  
A.(1) (2) (3)  
B.(2) (3) (4)  
C.(1) (2) (3) (4)  
D.(1) (2) (3) (4) (5)

关于OLAP和OLTP的区别描述,不正确的是: (C)  
A.OLAP主要是关于如何理解聚集的大量不同的数据.它与OTAP应用程序不同  
B.与OLAP应用程序不同,OLTP应用程序包含大量相对简单的事务  
C.OLAP的特点在于事务量大,但事务内容比较简单且重复率高  
D.OLAP是以数据仓库为基础的,但其最终数据来源与OLTP一样均来自底层的数据库系统,两者面对的用户是相同的

关于OLAP和OLTP的说法,下列不正确的是: (A)  
A.OLAP事务量大,但事务内容比较简单且重复率高  
B.OLAP的最终数据来源与OLTP不一样  
C.OLTP面对的是决策人员和高层管理人员  
D.OLTP以应用为核心,是应用驱动的

设X={1，2，3}是频繁项集，则可由X产生( C )个关联规则。  
A.4       B.5

C.6       D.7

考虑下面的频繁3-项集的集合：{1,2,3},{1,2,4},{1,2,5},{1,3,4},{1,3,5},{2,3,4},{2,3,5},{3,4,5}假定数据集中只有5个项,采用合并策略,由候选产生过程得到4-项集不包含（ C ）  
A.1,2,3,4     B.1,2,3,5

C.1,2,4,5     D.1,3,4,5

下面选项中t不是s的子序列的是  ( C )  
A.s=<{2,4},{3,5,6},{8}> t=<{2},{3,6},{8}>  
B.s=<{2,4},{3,5,6},{8}> t=<{2},{8}>  
C.s=<{1,2},{3,4}> t=<{1},{2}>  
D.s=<{2,4},{2,4}> t=<{2},{4}>

在图集合中发现一组公共子结构,这样的任务称为 ( B )  
A.频繁子集挖掘     B.频繁子图挖掘

C.频繁数据项挖掘   D.频繁模式挖掘

下列度量不具有反演性的是 (D)  
A.系数             B.几率

C.Cohen度量       D.兴趣因子

下列 ( A )不是将主观信息加入到模式发现任务中的方法。  
A.与同一时期其他数据对比  
B.可视化  
C.基于模板的方法  
D.主观兴趣度量

下面购物蓝能够提取的3-项集的最大数量是多少（C）

|  |  |
| --- | --- |
| TID | 项集 |
| 1 | 牛奶,啤酒,尿布 |
| 2 | 面包,黄油,牛奶 |
| 3 | 牛奶,尿布,饼干 |
| 4 | 面包,黄油,饼干 |
| 5 | 啤酒,饼干,尿布 |
| 6 | 牛奶,尿布,面包,黄油 |
| 7 | 面包,黄油,尿布 |
| 8 | 啤酒,尿布 |
| 9 | 牛奶,尿布,面包,黄油 |
| 10 | 啤酒,饼干 |

A.1     B.2     C.3     D.4

以下哪些算法是分类算法  （B）

A.DBSCAN   B.C4.5    C.K-Mean  D.EM

以下哪些分类方法可以较好地避免样本的不平衡问题（A）

A.KNN   B.SVM   C.Bayes   D.神经网络

决策树中不包含一下哪种结点 ( C )

根结点（root node)

内部结点（internal node）

外部结点（external node）

叶结点（leaf node）

以下哪项关于决策树的说法是错误的 (C)

A. 冗余属性不会对决策树的准确率造成不利的影响   
B. 子树可能在决策树中重复多次   
C. 决策树算法对于噪声的干扰非常敏感    
D. 寻找最佳决策树是NP完全问题

在基于规则分类器的中,依据规则质量的某种度量对规则排序,保证每一个测试记录都是由覆盖它的“最好的”规格来分类,这种方案称为 (B)  
A. 基于类的排序方案   
B. 基于规则的排序方案   
C. 基于度量的排序方案   
D. 基于规格的排序方案。

以下哪些算法是基于规则的分类器 (A)   
A.  C4.5         B. KNN

C. Naive Bayes    D.ANN

可用作数据挖掘分析中的关联规则算法有（C）。  
A. 决策树、对数回归、关联模式

B. K均值法、SOM神经网络

C. Apriori算法、FP-Tree算法

D. RBF神经网络、K均值法、决策树

如果对属性值的任一组合,R中都存在一条规则加以覆盖,则称规则集R中的规则为( B )

A.无序规则  B.穷举规则  C.互斥规则  D.有序规则

用于分类与回归应用的主要算法有: ( D )

A.Apriori算法、HotSpot算法

B.RBF神经网络、K均值法、决策树

C.K均值法、SOM神经网络

D.决策树、BP神经网络、贝叶斯

40)如果允许一条记录触发多条分类规则,把每条被触发规则的后件看作是对相应类的一次投票,然后计票确定测试记录的类标号,称为（A）    
    A.无序规则  B.穷举规则  C.互斥规则  D.有序规则

41)考虑两队之间的足球比赛：队0和队1。假设65%的比赛队0胜出,剩余的比赛队1获胜。队0获胜的比赛中只有30%是在队1的主场,而队1取胜的比赛中75%是主场获胜。如果下一场比赛在队1的主场进行队1获胜的概率为 (C)  
    A.0.75    B.0.35    C.0.4678    D.0.5738

42)以下关于人工神经网络（ANN）的描述错误的有 (A)  
A.神经网络对训练数据中的噪声非常鲁棒

B.可以处理冗余特征

C.训练ANN是一个很耗时的过程

D.至少含有一个隐藏层的多层神经网络

43)通过聚集多个分类器的预测来提高分类准确率的技术称为 (A)

A.组合(ensemble)     B.聚集(aggregate)

C.合并(combination)  D.投票(voting)

44)简单地将数据对象集划分成不重叠的子集,使得每个数据对象恰在一个子集中,这种聚类类型称作（ B ）

 A.层次聚类           B.划分聚类

 C.非互斥聚类         D.模糊聚类

45)在基本K均值算法里,当邻近度函数采用（ A ）的时候,合适的质心是簇中各点的中位数。  
     A.曼哈顿距离        B.平方欧几里德距离

C.余弦距离          D.Bregman散度    
46)（ C ）是一个观测值,它与其他观测值的差别如此之大,以至于怀疑它是由不同的机制产生的。  
     A.边界点            B.质心

C.离群点            D.核心点  
47)BIRCH是一种（ B ）。  
     A.分类器            B.聚类算法

C.关联分析算法      D.特征选择算法  
48)检测一元正态分布中的离群点,属于异常检测中的基于（ A ）的离群点检测。

  A.统计方法         B.邻近度

C.密度             D.聚类技术  
49)（ C ）将两个簇的邻近度定义为不同簇的所有点对的平均逐对邻近度，它是一种凝聚层次聚类技术。  
      A.MIN（单链）       B.MAX（全链）

C.组平均            D.Ward方法  
50)（ D ）将两个簇的邻近度定义为两个簇合并时导致的平方误差的增量,它是一种凝聚层次聚类技术。  
      A.MIN（单链）       B.MAX（全链）

C.组平均            D.Ward方法  
51) 下列算法中，不属于外推法的是（ B ）。  
      A.移动平均法        B.回归分析法

C.指数平滑法        D.季节指数法  
52) 关联规则的评价指标是：（ C ）。  
      A. 均方误差、均方根误差

B. Kappa统计、显著性检验

C. 支持度、置信度

D. 平均绝对误差、相对误差  
53)关于K均值和DBSCAN的比较,以下说法不正确的是（ A ）。  
   A.K均值丢弃被它识别为噪声的对象,而DBSCAN一般聚类所有对象。  
  B.K均值使用簇的基于原型的概念,而DBSCAN使用基于密度的概念。  
   C.K均值很难处理非球形的簇和不同大小的簇,DBSCAN可以处理不同大小和不同形状的簇。  
   D.K均值可以发现不是明显分离的簇,即便簇有重叠也可以发现,但是DBSCAN会合并有重叠的簇。

54）从研究现状上看，下面不属于云计算特点的是（ C  ）

A.超大规模           B.虚拟化

C.私有化             D.高可靠性

55)考虑这么一种情况：一个对象碰巧与另一个对象相对接近,但属于不同的类,因为这两个对象一般不会共享许多近邻,所以应该选择（ D ）的相似度计算方法。  
    A.平方欧几里德距离       B.余弦距离

C.直接相似度             D.共享最近邻  
56) 分析顾客消费行业，以便有针对性的向其推荐感兴趣的服务，属于（ A）问题。

A.关联规则挖掘           B.分类与回归

C.聚类分析               D.时序预测  
57)以下哪个聚类算法不是属于基于原型的聚类（ D ）。  
    A.模糊C均值         B.EM算法

C.SOM                D.CLIQUE  
58)关于混合模型聚类算法的优缺点,下面说法正确的是（ B ）。  
   A.当簇只包含少量数据点,或者数据点近似协线性时,混合模型也能很好地处理。  
   B.混合模型比K均值或模糊c均值更一般,因为它可以使用各种类型的分布。  
   C.混合模型很难发现不同大小和椭球形状的簇。  
   D.混合模型在有噪声和离群点时不会存在问题。  
59)以下哪个聚类算法不属于基于网格的聚类算法（ D ）。  
    A.STING         B.WaveCluster

C.MAFIA         D.BIRCH  
60)一个对象的离群点得分是该对象周围密度的逆。这是基于（ C ）的离群点定义。  
    A．概率          B．邻近度

C．密度          D．聚类

舆情研判，信息科学侧重（ C  ）,社会和管理科学侧重突发群体事件管理中的群体心理行为及舆论控制研究，新闻传播学侧重对舆论的本体进行规律性的探索和研究。

A.舆论的本体进行规律性的探索和研究

B.舆论控制研究

C.互联网文本挖掘和分析技术

D.用户行为分析

MapReduce的Map函数产生很多的（   C    ）

A.key                         B.value

C.<key,value>              D.Hash

Mapreduce适用于 （  D  ）

A.任意应用程序

B. 任意可在windows servet2008上运行的程序

C.可以串行处理的应用程序

D. 可以并行处理的应用程序

PageRank是一个函数,它对Web中的每个网页赋予一个实数值。它的意图在于网页的PageRank越高,那么它就（    D   ）。

A.相关性越高              B.越不重要

C.相关性越低              D.越重要

╳A.一对一                     B.一对多

C. 多对多                         D. 多对一

协同过滤分析用户兴趣,在用户群中找到指定用户的相似(兴趣)用户,综合这些用户对某一信息的评价,形成系统对该指定用户对此信息的喜好程度（   D    ）,并将这些用户喜欢的项推荐给有相似兴趣的用户。

A. 相似                   B.相同

C.推荐                      D. 预测

大数据指的是所涉及的资料量规模巨大到无法透过目前主流软件工具,在合理时间内达到撷取、管理、处理、并（   B    ）成为帮助企业经营决策更积极目的的信息。

A.收集                    B.整理

C.规划                      D.聚集

大数据科学关注大数据网络发展和运营过程中（  D    ）大数据的规律及其与自然和社会活动之间的关系。

A.大数据网络发展和运营过程             B.规划建设运营管理

C.规律和验证                                                        D.发现和验证

大数据的价值是通过数据共享、（    D    ）后获取最大的数据价值

A.算法共享                B.共享应用

C. 数据交换                D. 交叉复用

社交网络产生了海量用户以及实时和完整的数据,同时社交网络也记录了用户群体的（   C    ）,通过深入挖掘这些数据来了解用户,然后将这些分析后的数据信息推给需要的品牌商家或是微博营销公司。

A.地址                    B.行为

C.情绪                      D.来源

通过数据收集和展示数据背后的（    D    ）,运用丰富的、具有互动性的可视化手段,数据新闻学成为新闻学作为一门新的分支进入主流媒体,即用数据报道新闻。

A.数据收集                 B.数据挖掘

C.真相                      D. 关联与模式

CRISP-DM 模型中Evaluation表示对建立的模型进行评估,重点具体考虑得出的结果是否符合（   C    ）的商业目的。

A.第二步                 B.第三步

C.第一步                 D.最后一步

发现关联规则的算法通常要经过以下三个步骤：连接数据,作数据准备；给定最小支持度和（    D    ）,利用数据挖掘工具提供的算法发现关联规则；可视化显示、理解、评估关联规则

A. 最小兴趣度               B. 最小置信度

C. 最大支持度                 D. 最小可信度

规则I->j,“有可能”,等于所有包含I的购物篮中同时包含J的购物篮的比例,为（    B   ）。

A. 置信度                 B.可信度

C. 兴趣度                      D. 支持度

如果一个匹配中,任何一个节点都不同时是两条或多条边的端点,也称作（   C     ）

A. 极大匹配            B.二分匹配

C完美匹配               D.极小匹配

只要具有适当的政策推动,大数据的使用将成为未来提高竞争力、生产力、创新能力以及（   D   ）的关键要素。

A.提高消费              B.提高GDP

C.提高生活水平      D. 创造消费者盈余

个性化推荐系统是建立在海量数据挖掘基础上的一种高级商务智能平台,以帮助（    D  ）为其顾客购物提供完全个性化的决策支持和信息服务。

A.公司                 B.各单位

C.跨国企业               D. 电子商务网站

云计算是对（  D  ）技术的发展与运用

A.并行计算        B.网格计算

C.分布式计算      D.三个选项都是

（ B  ）是Google提出的用于处理海量数据的并行编程模式和大规模数据集的并行运算的软件架构。

A.GFS            B.MapReduce

C.Chubby         D.BitTable

在Bigtable中，（ A ）主要用来存储子表数据以及一些日志文件

A. GFS           B. Chubby

C.SSTable         D.MapReduce

二、判断题（共40题）

分类是预测数据对象的离散类别，预测是用于数据对象的连续取值。  (对)

时序预测回归预测一样，也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。(错)

数据挖掘的主要任务是从数据中发现潜在的规则，从而能更好的完成描述数据、预测数据等任务。 (对)

对遗漏数据的处理方法主要有：忽略该条记录；手工填补遗漏值；利用默认值填补遗漏值；利用均值填补遗漏值；利用同类别均值填补遗漏值；利用最可能的值填充遗漏值。(对)

神经网络对噪音数据具有高承受能力，并能对未经过训练的数据具有分类能力，但其需要很长的训练时间，因而对于有足够长训练时间的应用更合适。(对)

数据分类由两步过程组成：第一步，建立一个聚类模型，描述指定的数据类集或概念集；第二步，使用模型进行分类。(错)

聚类是指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。 (对)

决策树方法通常用于关联规则挖掘。 (错)

数据规范化指将数据按比例缩放(如更换大单位)，使之落入一个特定的区域（如0-1）以提高数据挖掘效率的方法。规范化的常用方法有：最大-最小规范化、零-均值规范化、小数定标规范化。(对)

原始业务数据来自多个数据库或数据仓库，它们的结构和规则可能是不同的，这将导致原始数据非常的杂乱、不可用，即使在同一个数据库中，也可能存在重复的和不完整的数据信息，为了使这些数据能够符合数据挖掘的要求，提高效率和得到清晰的结果，必须进行数据的预处理。(对)

数据取样时，除了要求抽样时严把质量关外，还要求抽样数据必须在足够范围内有代表性。(对)

分类规则的挖掘方法通常有：决策树法、贝叶斯法、人工神经网络法、粗糙集法和遗传算法。(对)

可信度是对关联规则的准确度的衡量。  (错)

孤立点在数据挖掘时总是被视为异常、无用数据而丢弃。  (错)

Apriori算法是一种典型的关联规则挖掘算法。(对)

用于分类的离散化方法之间的根本区别在于是否使用类信息。　（对）

特征提取技术并不依赖于特定的领域。　　（错）

模型的具体化就是预测公式，公式可以产生与观察值有相似结构的输出，这就是预测值。　（对）

文本挖掘又称信息检索，是从大量文本数据中提取以前未知的、有用的、可理解的、可操作的知识的过程。（错）

定量属性可以是整数值或者是连续值。　　（对）

可视化技术对于分析的数据类型通常不是专用性的。　（错）

OLAP技术侧重于把数据库中的数据进行分析、转换成辅助决策信息，是继数据库技术发展之后迅猛发展起来的一种新技术。（对）

Web数据挖掘是通过数据库仲的一些属性来预测另一个属性，它在验证用户提出的假设过程中提取信息。   （错）

关联规则挖掘过程是发现满足最小支持度的所有项集代表的规则。  （错）

利用先验原理可以帮助减少频繁项集产生时需要探查的候选项个数。  （对）

先验原理可以表述为：如果一个项集是频繁的，那包含它的所有项集也是频繁的。（错）

回归分析通常用于挖掘关联规则。（错）

具有较高的支持度的项集具有较高的置信度。（错）

维归约可以去掉不重要的属性，减少数据立方体的维数，从而减少数据挖掘处理的数据量，提高挖掘效率。（对）

聚类（clustering）是这样的过程：它找出描述并区分数据类或概念的模型(或函数)，以便能够使用模型预测类标记未知的对象类。  （错）

对于SVM分类算法，待分样本集中的大部分样本不是支持向量，移去或者减少这些样本对分类结果没有影响。（对）

Bayes法是一种在已知后验概率与类条件概率的情况下的模式分类方法，待分样本的分类结果取决于各类域中样本的全体。   (错)

在决策树中，随着树中结点数变得太大，即使模型的训练误差还在继续减低，但是检验误差开始增大，这是出现了模型拟合不足的问题。（错）

在聚类分析当中，簇内的相似性越大，簇间的差别越大，聚类的效果就越差。（错）

聚类分析可以看作是一种非监督的分类。（对）

K均值是一种产生划分聚类的基于密度的聚类算法，簇的个数由算法自动地确定。（错

基于邻近度的离群点检测方法不能处理具有不同密度区域的数据集。（对）

如果一个对象不强属于任何簇，那么该对象是基于聚类的离群点。（对）

大数据的4V特点是Volume、Velocity、Variety、Veracity。（对）

聚类分析的相异度矩阵是用于存储所有对象两两之间相异度的矩阵，为一个nn维的单模矩阵。（对）

三、多选题（共30题）

噪声数据的产生原因主要有：（ABCD）

数据采集设备有问题

在数据录入过程中发生了人为或计算机错误

数据传输过程中发生错误

由于命名规则或数据代码不同而引起的不一致

寻找数据集中的关系是为了寻找精确、方便并且有价值地总结出数据的某一特征的表示,这个过程包括了以下哪些步骤？ (A B C D)

A.选择一个算法过程使评分函数最优  
B.决定如何量化和比较不同表示拟合数据的好坏  
C.决定要使用的表示的特征和结构  
D.决定用什么样的数据管理原则以高效地实现算法

数据挖掘的预测建模任务主要包括哪几大类问题？(A B)

A. 分类

B. 回归

C. 聚类

D. 关联规则挖掘

4) 下列属于不同的有序数据的有：(A B C D)  
A.时序数据

B.序列数据

C.时间序列数据

D.事务数据

E.空间数据

5) 下面属于数据集的一般特性的有：( B C D)  
A.连续性

B.维度

C.稀疏性

D.分辨率

E.相异性

6）下面属于维归约常用的处理技术的有： (A C)  
A.主成分分析

B.特征提取

C.奇异值分解

D.特征加权

E.离散化

噪声数据处理的方法主要有：（ ABD ）

A.分箱           B.聚类

C.关联分析       D.回归

数据挖掘的主要功能包括概念描述、趋势分析、孤立点分析及（  ABCD  ）等方面。

A.挖掘频繁模式  　   B.分类和预测

C.聚类分析           D.偏差分析

以下各项均是针对数据仓库的不同说法,你认为正确的有（BCD    ）。  
  A．数据仓库就是数据库  
  B．数据仓库是一切商业智能系统的基础  
  C．数据仓库是面向业务的,支持联机事务处理（OLTP）  
  D．数据仓库支持决策而非事务处理

10）联机分析处理包括(  BCD  )基本分析功能。  
A.聚类             B.切片

C.转轴              D.切块

11）利用Apriori算法计算频繁项集可以有效降低计算频繁集的时间复杂度。在以下的购物篮中产生支持度不小于3的候选3-项集，在候选2-项集中需要剪枝的是( BD )

|  |  |
| --- | --- |
| TID | 项集 |
| 1 | 面包、牛奶 |
| 2 | 面包、尿布、啤酒、鸡蛋 |
| 3 | 牛奶、尿布、啤酒、可乐 |
| 4 | 面包、牛奶、尿布、啤酒 |
| 5 | 面包、牛奶、尿布、可乐 |

A.啤酒、尿布        B.啤酒、面包

C.面包、尿布        D.啤酒、牛奶

12）下表是一个购物篮,假定支持度阈值为40%,其中(  AD   )是频繁闭项集。

|  |  |
| --- | --- |
| ID | 项集 |
| 1 | 面包、牛奶、尿布 |
| 2 | 面包、牛奶、尿布、啤酒 |
| 3 | 牛奶、尿布、鸡蛋 |
| 4 | 面包、尿布、啤酒、鸡蛋 |
| 5 | 啤酒、鸡蛋 |

A.面包、牛奶、尿布      B.面包、啤酒  
C.尿布、啤酒            D.啤酒、鸡蛋

13）Apriori算法的计算复杂度受(  ABCD  )影响。

A.支持度阀值            B.项数（维度）

C.事务数                D.事务平均宽度

14）以下关于非频繁模式说法,正确的是（  AD  ）

A.其支持度小于阈值         B.都是不让人感兴趣的

C.包含负模式和负相关模式   D.对异常数据项敏感

15）以下属于分类器评价或比较尺度的有: (   ACD   )

A.预测准确度           B.召回率

C.模型描述的简洁度     D.计算复杂度

16）贝叶斯信念网络(BBN)有如下哪些特点。 （AB）

A.构造网络费时费力

B.对模型的过分问题非常鲁棒

C.贝叶斯网络不适合处理不完整的数据

D.网络结构确定后,添加变量相当麻烦

17）如下哪些不是最近邻分类器的特点。  (C)

A.它使用具体的训练实例进行预测，不必维护源自数据的模型

B.分类一个测试样例开销很大

C.最近邻分类器基于全局信息进行预测

D.可以生产任意形状的决策边界

18）以下属于聚类算法的是（ AB ）。  
    A.K-Means         B.DBSCAN

C.Apriori         D.KNN

19)（ CD ）都属于簇有效性的监督度量。  
    A.轮廓系数        B.共性分类相关系数

C.熵              D.F度量

20)（ ABCD ）这些数据特性都是对聚类分析具有很强影响的。  
    A.高维性          B.规模

C.稀疏性          D.噪声和离群点

21)在聚类分析当中，（ AD ）等技术可以处理任意形状的簇。  
    A.MIN（单链）     B.MAX（全链）

C.组平均          D.Chameleon

22)（ AB ）都属于分裂的层次聚类算法。  
    A.二分K均值      B.MST

C.Chameleon       D.组平均

23)数据挖掘的挖掘方法包括：（   ABCD    ）

A.聚类分析        B.回归分析

C.神经网络               D.决策树算法

24)Web内容挖掘实现技术（   ABCD   ）

A.文本总结        B.文本分类

C.文本聚类        D.关联规则

25)基于内容的推荐生成推荐的过程主要依靠（  ACD    ）

A.内容分析器

B.推荐系统

C.文件学习器

D.过滤部件

26)云计算的服务方式有（ ACD  ）

A.IaaS              B.Raas

C.PaaS              D.SaaS

27)文本挖掘的工具有（ BCD  ）

A.SPP Text Mining

B.IBM DB2 intelligent Miner

C.SAS Text Miner

D.SPSS Text Mining

28)推荐系统为客户推荐商品,自动完成个性化选择商品的过程,满足客户的个性化需求,推荐基于网站最热卖商品、客户所处城市、（    D   ）,推测客户将来可能的购买行为。

A.客户的朋友            B.客户的个人信息

C.客户的兴趣爱好    D.客户过去的购买行为和购买记录

29)数据预处理方法主要有（ ABCD）。

A.数据清洗        B.数据集成

C.数据变换        D.数据归约

30)与传统的分布式程序设计相比，Mapreduce封装了  （  ABCD  ）等细节，还提供了一个简单而强大的接口。

A.并行处理        B.容错处理

C.本地化计算      D.负载均衡