1. 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

2. 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  
C. Precision, ROC  
D. Recall, ROC

3. 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

4. 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

5. 什么是KDD？ (A)  
A. 数据挖掘与知识发现  
B. 领域知识发现  
C. 文档知识发现  
D. 动态知识发现

6. 使用交互式的和可视化的技术，对数据进行探索属于数据挖掘的哪一类任务？（A）  
A. 探索性数据分析  
B. 建模描述  
C. 预测建模  
D. 寻找模式和规则

7. 为数据的总体分布建模；把多维空间划分成组等问题属于数据挖掘的哪一类任务？(B)  
A. 探索性数据分析  
B. 建模描述  
C. 预测建模  
D. 寻找模式和规则

8. 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

9. 用户有一种感兴趣的模式并且希望在数据集中找到相似的模式，属于数据挖掘哪一类任务？(A)  
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

11.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

12. 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

13.上题中，等宽划分时（宽度为50），15又在哪个箱子里？ (A)  
A 第一个  
B 第二个  
C 第三个  
D 第四个

14.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

15. 在上题中，属于定量的属性类型是：(C)  
A 标称  
B 序数  
C 区间  
D 相异

16. 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

17. 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

18.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

19. 考虑值集{1、2、3、4、5、90}，其截断均值（p=20%）是 (C)  
A 2  
B 3  
C 3.5  
D 5

20. 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

21. 熵是为消除不确定性所需要获得的信息量，投掷均匀正六面体骰子的熵是： (B)  
A 1比特  
B 2.6比特  
C 3.2比特  
D 3.8比特

22. 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

23.假定用于分析的数据包含属性age。数据元组中age的值如下（按递增序）：13，15，16，16，19，20，20，21，22，22，25，25，25，30，33，33，35，35，36，40，45，46，52，70, 问题：使用按箱平均值平滑方法对上述数据进行平滑，箱的深度为3。第二个箱子  
值为：(A)  
A 18.3  
B 22.6  
C 26.8  
D 27.9

24. 考虑值集{12 24 33 2 4 55 68 26}，其四分位数极差是：(A)  
A 31  
B 24  
C 55  
D 3

25. 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

26. 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

27. 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

28. 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.

29. 关于基本数据的元数据是指: (D)  
A. 基本元数据与数据源,数据仓库,数据集市和应用程序等结构相关的信息;  
B. 基本元数据包括与企业相关的管理方面的数据和信息;  
C. 基本元数据包括日志文件和简历执行处理的时序调度信息;  
D. 基本元数据包括关于装载和更新处理,分析处理以及管理方面的信息.

30. 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.

31. 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

32. 在有关数据仓库测试,下列说法不正确的是: (D)  
A. 在完成数据仓库的实施过程中,需要对数据仓库进行各种测试.测试工作中要包括单元测试和系统测试.  
B. 当数据仓库的每个单独组件完成后,就需要对他们进行单元测试.  
C. 系统的集成测试需要对数据仓库的所有组件进行大量的功能测试和回归测试.  
D. 在测试之前没必要制定详细的测试计划.

33. OLAP技术的核心是: (D)  
A. 在线性;  
B. 对用户的快速响应;  
C. 互操作性.  
D. 多维分析;

34. 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)

35. 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.

36. OLAM技术一般简称为”数据联机分析挖掘”,下面说法正确的是: (D)  
A. OLAP和OLAM都基于客户机/服务器模式,只有后者有与用户的交互性;  
B. 由于OLAM的立方体和用于OLAP的立方体有本质的区别.  
C. 基于WEB的OLAM是WEB技术与OLAM技术的结合.  
D. OLAM服务器通过用户图形借口接收用户的分析指令,在元数据的知道下,对超级立方体作一定的操作.

37. 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.

38. 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

40. 概念分层图是\_\_(B)\_\_图。  
A、无向无环  
B、有向无环  
C、有向有环  
D、无向有环

41. 频繁项集、频繁闭项集、最大频繁项集之间的关系是： (C)  
A、频繁项集 频繁闭项集 =最大频繁项集  
B、频繁项集 = 频繁闭项集 最大频繁项集  
C、频繁项集 频繁闭项集 最大频繁项集  
D、频繁项集 = 频繁闭项集 = 最大频繁项集

42. 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

43.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}>

44. 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

45. 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

46. 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

47. 下面购物篮能够提取的3-项集的最大数量是多少（C）

ID 购买项  
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

48. 以下哪些算法是分类算法，（B）  
A，DBSCAN  
B，C4.5  
C,K-Mean  
D,EM

49. 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

50. 决策树中不包含一下哪种结点， (C)

A,根结点（root node)  
B,内部结点（internal node）  
C,外部结点（external node）  
D,叶结点（leaf node）

51. 不纯性度量中Gini计算公式为（其中c是类的个数） (A)  
A, B, C, D, （A）

53. 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

54. 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。

55. 以下哪些算法是基于规则的分类器 (A)  
A. C4.5  
B. KNN  
C. Na?ve Bayes  
D. ANN

56. 如果规则集R中不存在两条规则被同一条记录触发，则称规则集R中的规则为（C）；  
A, 无序规则  
B，穷举规则  
C， 互斥规则  
D，有序规则

57. 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

58. 如果规则集中的规则按照优先级降序排列，则称规则集是 (D)  
A, 无序规则  
B，穷举规则  
C， 互斥规则  
D，有序规则

59. 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

60. 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

61. 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

62. 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)

63. 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

64. 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

65.（ 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

66. 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

67. 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

68.（ 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

69.（ 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

70. DBSCAN在最坏情况下的时间复杂度是（ B ）。  
A、O(m)  
B、O(m2)  
C、O(log m)  
D、O(m\*log m)

71. 在基于图的簇评估度量表里面，如果簇度量为proximity(Ci , C)，簇权值为mi ，那么它的类型是（ C ）。  
A、基于图的凝聚度  
B、基于原型的凝聚度  
C、基于原型的分离度  
D、基于图的凝聚度和分离度

72. 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。

73. 以下是哪一个聚类算法的算法流程：①构造k－最近邻图。②使用多层图划分算法划分图。③repeat：合并关于相对互连性和相对接近性而言，最好地保持簇的自相似性的簇。④until：不再有可以合并的簇。（ C ）。  
A、MST  
B、OPOSSUM  
C、Chameleon  
D、Jarvis－Patrick（JP）

74. 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

75. 以下属于可伸缩聚类算法的是（ A ）。  
A、CURE  
B、DENCLUE  
C、CLIQUE  
D、OPOSSUM

76. 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

77. 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。

78. 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

79. 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

80. 下面关于Jarvis－Patrick（JP）聚类算法的说法不正确的是（ D ）。  
A、JP聚类擅长处理噪声和离群点，并且能够处理不同大小、形状和密度的簇。  
B、JP算法对高维数据效果良好，尤其擅长发现强相关对象的紧致簇。  
C、JP聚类是基于SNN相似度的概念。  
D、JP聚类的基本时间复杂度为O(m)。

**二、 多选题**

1. 通过数据挖掘过程所推倒出的关系和摘要经常被称为：(A B)  
A. 模型  
B. 模式  
C. 模范  
D. 模具

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

3. 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. 模式发现  
D. 模式匹配

4. 数据挖掘算法的组件包括：(A B C D)  
A. 模型或模型结构  
B. 评分函数  
C. 优化和搜索方法  
D. 数据管理策略

5. 以下哪些学科和数据挖掘有密切联系？(A D)  
A. 统计  
B. 计算机组成原理  
C. 矿产挖掘  
D. 人工智能

6. 在现实世界的数据中，元组在某些属性上缺少值是常有的。描述处理该问题的各种方法有： (ＡＢＣＤＥ)  
A忽略元组  
B使用属性的平均值填充空缺值  
C使用一个全局常量填充空缺值  
D使用与给定元组属同一类的所有样本的平均值  
E使用最可能的值填充空缺值

7.下面哪些属于可视化高维数据技术 (ＡＢＣＥ)  
A 矩阵  
B 平行坐标系  
C星形坐标 D散布图  
E Chernoff脸

8. 对于数据挖掘中的原始数据，存在的问题有： (ＡＢＣＤＥ)  
A 不一致  
B重复  
C不完整  
D 含噪声  
E 维度高

9.下列属于不同的有序数据的有：(ＡＢＣＥ)  
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

10.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

11. 下面属于维归约常用的线性代数技术的有： (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

12. 下面列出的条目中，哪些是数据仓库的基本特征： (ACD)  
A. 数据仓库是面向主题的  
B. 数据仓库的数据是集成的  
C. 数据仓库的数据是相对稳定的  
D. 数据仓库的数据是反映历史变化的  
E. 数据仓库是面向事务的

13. 以下各项均是针对数据仓库的不同说法，你认为正确的有（BCDE ）。  
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  
E．数据仓库的主要目标就是帮助分析，做长期性的战略制定

14. 数据仓库在技术上的工作过程是： (ABCD)  
A. 数据的抽取  
B. 存储和管理  
C. 数据的表现  
D. 数据仓库设计  
E. 数据的表现

15. 联机分析处理包括以下哪些基本分析功能？ (BCD)  
A. 聚类  
B. 切片  
C. 转轴  
D. 切块  
E. 分类

16. 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）  
ID 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

17. 下表是一个购物篮，假定支持度阈值为40%，其中\_\_(A D)\_\_是频繁闭项集。  
TID 项  
1 abc  
2 abcd  
3 bce  
4 acde  
5 de  
A、abc  
B、ad  
C、cd  
D、de

18. 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

19. 非频繁模式\_\_(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

20. 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

21. 在评价不平衡类问题分类的度量方法有如下几种，(ABCD)  
A,F1度量  
B,召回率（recall）  
C,精度（precision）  
D,真正率（ture positive rate,TPR）

22. 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

23. 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

24. 如下那些不是基于规则分类器的特点，（AC）  
A,规则集的表达能力远不如决策树好  
B,基于规则的分类器都对属性空间进行直线划分，并将类指派到每个划分  
C,无法被用来产生更易于解释的描述性模型  
D,非常适合处理类分布不平衡的数据集

25. 以下属于聚类算法的是（ ABD ）。  
A、K均值  
B、DBSCAN  
C、Apriori  
D、Jarvis-Patrick（JP）

26.（ 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

27. 簇有效性的面向相似性的度量包括（ BC ）。  
A、精度  
B、Rand统计量  
C、Jaccard系数  
D、召回率

28.（ 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

29. 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

30. （ 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

**三、 判断题**

1. 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)

2. 数据挖掘的目标不在于数据采集策略，而在于对于已经存在的数据进行模式的发掘。（对）3. 图挖掘技术在社会网络分析中扮演了重要的角色。（对）

4. 模式为对数据集的全局性总结，它对整个测量空间的每一点做出描述；模型则对变量变化空间的一个有限区域做出描述。（错）

5. 寻找模式和规则主要是对数据进行干扰，使其符合某种规则以及模式。（错）

6. 离群点可以是合法的数据对象或者值。　　　　（对）

7. 离散属性总是具有有限个值。　　　　　　　　（错）

8. 噪声和伪像是数据错误这一相同表述的两种叫法。　　　　　（错）

9. 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）

10. 特征提取技术并不依赖于特定的领域。　　　　　　（错）

11. 序列数据没有时间戳。　　　　　　（对）

12. 定量属性可以是整数值或者是连续值。　　　　　（对）

13. 可视化技术对于分析的数据类型通常不是专用性的。　　　　（错）

14. DSS主要是基于数据仓库.联机数据分析和数据挖掘技术的应用。（对）

15. 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）

16. 商业智能系统与一般交易系统之间在系统设计上的主要区别在于：后者把结构强加于商务之上，一旦系统设计完毕，其程序和规则不会轻易改变；而前者则是一个学习型系统，能自动适应商务不断变化的要求。 （对）

17. 数据仓库中间层OLAP服务器只能采用关系型OLAP （错）

18．数据仓库系统的组成部分包括数据仓库，仓库管理，数据抽取，分析工具等四个部分. (错)

19. 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）

21. 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）

22. 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）。

23. 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）

24. 如果规则 不满足置信度阈值，则形如 的规则一定也不满足置信度阈值，其中 是X的子集。（对）

25. 具有较高的支持度的项集具有较高的置信度。（错）

26. 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）

27. 分类和回归都可用于预测，分类的输出是离散的类别值，而回归的输出是连续数值。(对)

28. 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）

29. 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)

30.分类模型的误差大致分为两种：训练误差（training error）和泛化误差（generalization error）. (对)

31. 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）

32. SVM是这样一个分类器，他寻找具有最小边缘的超平面，因此它也经常被称为最小边缘分类器（minimal margin classifier） (错)

33. 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）

34. 聚类分析可以看作是一种非监督的分类。（对）

35. 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）

36. 给定由两次运行K均值产生的两个不同的簇集，误差的平方和最大的那个应该被视为较优。（错）

37. 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）

38. 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）

39. 从点作为个体簇开始，每一步合并两个最接近的簇，这是一种分裂的层次聚类方法。（错）40. DBSCAN是相对抗噪声的，并且能够处理任意形状和大小的簇。（对）