

《海量数据挖掘技术及工程实践》题目

一、单选题（共 80 题）

- 1) (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
- 2) 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
- 3) 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
- 4) 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
- 5) 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
- 6) 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
- 7) 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

- 8) 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
- 9) 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
- 10) 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
- 11) 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
- 12) 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
- 13) 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
- 14) 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)
- 15) 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。则年纪属性的众数是: (A)
- A. 一年级 B. 二年级
C. 三年级 D. 四年级
- 16) 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

17) 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

18) 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

19) 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

20) 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

21) 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)

22) 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

23) 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

24) 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

25) 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

26) 下面选项中 t 不是 s 的子序列的是 (C)

A. $s=\langle\{2,4\},\{3,5,6\},\{8\}\rangle$ $t=\langle\{2\},\{3,6\},\{8\}\rangle$

B. $s=\langle\{2,4\},\{3,5,6\},\{8\}\rangle$ $t=\langle\{2\},\{8\}\rangle$

C. $s=\langle\{1,2\},\{3,4\}\rangle$ $t=\langle\{1\},\{2\}\rangle$

D. $s=\langle\{2,4\},\{2,4\}\rangle$ $t=\langle\{2\},\{4\}\rangle$

27) 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

28) 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

29) 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

30) 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

31) 以下哪些算法是分类算法 (B)

A.DBSCAN B.C4.5 C.K-Mean D.EM

32) 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

33) 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)

- 34) 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
- 35) 在基于规则分类器中,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。
- 36) 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
- 37) 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
- 38) 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
- 39) 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)

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 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 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

61) 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

62) MapReduce的 Map函数产生很多的(C)

- A.key B.value
- C.<key,value> D.Hash

63) 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
- 64) 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
- 65) 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
- 66) 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
- 67) 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
- 68) 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
- 69) 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

- 70) 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
- 71) 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
- 72) 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
- 73) 规则 $I \rightarrow j$, “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
- 74) 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
- 75) 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
- 76) 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
- 77) 云计算是对 (D) 技术的发展与运用

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/248026116062006100>