一、 填空题

1. Web挖掘可分为 ____、____和____3 大类。
2. 数据仓库需要统一数据源， 包括统一 ____、统一 ____、统一 ____和统一数据特征 4 个方面。
3. 数据分割通常按时间、 ____、 ____、 ____以及组合方法进行。
4. 噪声数据处理的方法主要有 ____、____和____。
5. 数值归约的常用方法有 ____、____、____、____和对数模型等。
6. 评价关联规则的 2 个主要指标是 和 。
7. 多维数据集通常采用 ____或雪花型架构，以 ____表为中心，连接多个 ____表 。
8. 决策树是用 ____作为结点，用 ____作为分支的树结构。
9. 关联可分为简单关联、 ____和____。
10. BP 神经网络的作用函数通常为 ____区间的____。
11. 数据挖掘的过程主要包括确定业务对象、 ____、____、____及知识同化等几个步骤。
12. 数据挖掘技术主要涉及 ____、____和____3 个技术领域。
13. 数据挖掘的主要功能包括 ____、____、____、____、趋势分析、孤立点分析和偏差分析 7 个方面。
14. 人工神经网络具有 ____和____等特点，其结构模型包括 ____、____和自组织网络 3 种。
15. 数据仓库数据的 4 个基本特征是 ____、____、非易失、随时间变化。
16. 数据仓库的数据通常划分为 ____、____、____和____等几个级别。
17. 数据预处理的主要内容（方法）包括____、____、____和 数据归约等。
18. 平滑分箱数据的方法主要有 ____、____和____。
19. 数据挖掘发现知识的类型主要有广义知识、 ____、____、____和偏差型知识五种。
20. OLAP的数据组织方式主要有 ____和____两种。
21. 常见的 OLAP多维数据分析包括 ____、____、____和旋转等操作。
22. 传统的决策支持系统是以 ____和____驱动，而新决策支持系统则是以 ____、建立在 ____和____技术之上。
23. OLAP的数据组织方式主要有 ____和____2 种。
24. SQL Server2000 的 OLAP组件叫 ____，OLAP操作窗口叫 ____。
25. BP 神经网络由 ____、____以及一或多个 ____结点组成。
26. 遗传算法包括 ____、____、____3 个基本算子。
27. 聚类分析的数据通常可分为区间标度变量、 ____、____、____、序数型以及混合类型等。
28. 聚类分析中最常用的距离计算公式有 ____、____、____等。
29. 基于划分的聚类算法有 ____和____。
30. Clementine 的工作流通常由 ____、____和____等节点连接而成。

31. 简单地说，数据挖掘就是从 ____中挖掘 ____的过程。

32. 数据挖掘相关的名称还有 ____、____、____等。

二、 判断题

( )1. 数据仓库的数据量越大，其应用价值也越大。

( )2. 啤酒与尿布的故事是聚类分析的典型实例。

( )3. 等深分箱法使每个箱子的记录个数相同。

( )4. 数据仓库“粒度”越细，记录数越少。

( )5. 数据立方体由 3 维构成， Z 轴表示事实数据。

( )6. 决策树方法通常用于关联规则挖掘。

( )7. ID3 算法是决策树方法的早期代表。

( )8. C4.5 是一种典型的关联规则挖掘算法。

( )9. 回归分析通常用于挖掘关联规则。

( )10. 人工神经网络特别适合解决多参数大复杂度问题。

( )11. 概念关系分析是文本挖掘所独有的。

( )12. 可信度是对关联规则的准确度的衡量。

( )13. 孤立点在数据挖掘时总是被视为异常、无用数据而丢弃。

( )14. SQL Server 2000 不提供关联规则挖掘算法。

( )15. Clementine 是 IBM 公司的专业级数据挖掘软件。

( )16. 决策树方法特别适合于处理数值型数据。

( )17. 数据仓库的数据为历史数据，从来不需要更新。

( )18. 等宽分箱法使每个箱子的取值区间相同。

( )19. 数据立方体是广义知识发现的方法和技术之一。

( )20. 数据立方体的其中一维用于记录事实数据。

( )21. 决策树通常用于分类与预测。

( )22. Apriori 算法是一种典型的关联规则挖掘算法。

( )23. 支持度是衡量关联规则重要性的一个指标。

( )24. SQL Server 2000 集成了 OLAP，但不具有数据挖掘功能。

( )25. 人工神经网络常用于分类与预测。

三、 名词解释

1. 数据仓库： 是一种新的数据处理体系结构，是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化 (不同时间) 的数据集合，为企业决策支持系统提供所需的集成信息。
2. 孤立点： 指数据库中包含的一些与数据的一般行为或模型不一致的异常数据。
3. OLAP: OLAP是在 OLTP的基础上发展起来的，以数据仓库为基础的数据分析处理，是共享多维信息的快速分析，是被专门设计用于支持复杂的分析操作，侧

重对分析人员和高层管理人员的决策支持。

4. 粒度： 指数据仓库的数据单位中保存数据细化或综合程度的级别。粒度影响存放在数据仓库中的数据量的大小，同时影响数据仓库所能回答查询问题的细节程度。
5. 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。
6. 关联知识： 是反映一个事件和其他事件之间依赖或相互关联的知识。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。
7. 数据挖掘： 从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
8. OLTP: OLTP为联机事务处理的缩写， OLAP是联机分析处理的缩写。 前者是以数据库为基础的， 面对的是操作人员和低层管理人员， 对基本数据进行查询和增、删、改等处理。
9. ROLAP: 是基于关系数据库存储方式的， 在这种结构中， 多维数据被映像成二维关系表，通常采用星型或雪花型架构，由一个事实表和多个维度表构成。
10. MOLAP: 是基于类似于“ 超立方 ”块的 OLAP存储结构，由许多经压缩的、类似于多维数组的对象构成，并带有高度压缩的索引及指针结构，通过直接偏移计算进行存取。
11. 数据归约： 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。
12. 广义知识： 通过对大量数据的归纳、概括和抽象，提炼出带有普遍性的、概括性的描述统计的知识。
13. 预测型知识： 是根据时间序列型数据，由历史的和当前的数据去推测未来的数据，也可以认为是以时间为关键属性的关联知识。
14. 偏差型知识：是对差异和极端特例的描述， 用于揭示事物偏离常规的异常现象，如标准类外的特例，数据聚类外的离群值等。
15. 遗传算法： 是一种优化搜索算法，它首先产生一个初始可行解群体，然后对这个群体通过模拟生物进化的选择、交叉、变异等遗传操作遗传到下一代群体，并最终达到全局最优。
16. 聚类：是将物理或抽象对象的集合分组成为多个类或簇 (cluster) 的过程， 使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。
17. 决策树： 是用样本的属性作为结点，用属性的取值作为分支的树结构。它是分类规则挖掘的典型方法，可用于对新样本进行分类。
18. 相异度矩阵： 是聚类分析中用于表示各对象之间相异度的一种矩阵， n 个对象的相异度矩阵是一个 nn 维的单模矩阵，其对角线元素均为 0，对角线两侧元素

的值相同。

19. 频繁项集： 指满足最小支持度的项集，是挖掘关联规则的基本条件之一。
20. 支持度： 规则 A→B 的支持度指的是所有事件中　A 与 B 同地发生的的概率，即 P(A∪B)，是 AB同时发生的次数与事件总次数之比。支持度是对关联规则重要性的衡量。
21. 可信度：规则 A→B 的可信度指的是包含　A项集的同时也包含　B 项集的条件概率 P(B|A)，是 AB同时发生的次数与　A 发生的所有次数之比。 可信度是对关联规则的准确度的衡量。
22. 关联规则：同时满足最小支持度阈值和最小可信度阈值的规则称之为关联规则。

四、 综合题

1. 何谓数据挖掘？它有哪些方面的功能？

从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、 但又是潜在有用的信息和知识的过程称为数据挖掘。相关的名称有知识发现、数据分析、数据融合、决策支持等。

数据挖掘的功能包括：概念描述、关联分析、分类与预测、聚类分析、趋势分析、孤立点分析以及偏差分析等。

2. 何谓数据仓库？为什么要建立数据仓库？

数据仓库是一种新的数据处理体系结构，是面向主题的、集成的、　不可更新的(稳定性)、随时间不断变化 (不同时间) 的数据集合， 为企业决策支持系统提供所需的集成信息。

建立数据仓库的目的有　3 个：

一是为了解决企业决策分析中的系统响应问题，　数据仓库能提供比传统事务数据库更快的大规模决策分析的响应速度。

二是解决决策分析对数据的特殊需求问题。　决策分析需要全面的、 正确的集成数据，这是传统事务数据库不能直接提供的。

三是解决决策分析对数据的特殊操作要求。　决策分析是面向专业用户而非一般业务员， 需要使用专业的分析工具，　对分析结果还要以商业智能的方式进行表现，这是事务数据库不能提供的。

3. 列举操作型数据与分析型数据的主要区别。

操作型数据	分析型数据
当前的、细节的	历史的、综合的
面向应用、事务驱动	面向分析、分析驱动
频繁增、删、改	几乎不更新，定期追加
操作需求事先知道	分析需求事先不知道
生命周期符合　SDLC	完全不同的生命周期
对性能要求高	对性能要求宽松

一次操作数据量小	一次操作数据量大
支持日常事务操作	支持管理决策需求

4. 何谓 OLTP和 OLAP? 它们的主要异同有哪些？

OLTP即联机事务处理，是以传统数据库为基础、面向操作人员和低层管理人员、对基本数据进行查询和增、删、改等的日常事务处理。 OLAP即联机分析处理，是在 OLTP基础上发展起来的、以数据仓库基础上的、面向高层管理人员和专业分析人员、为企业决策支持服务。

OLTP和 OLAP的主要区别如下表：

OLTP	OLAP
数据库数据	数据库或数据仓库数据
细节性数据	综合性数据
当前数据	历史数据
经常更新	不更新，但周期性刷新
一次性处理的数据量小	一次处理的数据量大
对响应时间要求高	响应时间合理
用户数量大	用户数据相对较少
面向操作人员，支持日常操作	面向决策人员， 支持管理需要
面向应用，事务驱动	面向分析，分析驱动

5. 何谓粒度？它对数据仓库有什么影响？按粒度组织数据的方式有哪些？

粒度是指数据仓库的数据单位中保存数据细化或综合程度的级别。 粒度影响存放在数据仓库中的数据量的大小， 同时影响数据仓库所能回答查询问题的细节程度。按粒度组织数据的方式主要有：

① 简单堆积结构

② 轮转综合结构

③ 简单直接结构

④ 连续结构

6. 简述数据仓库设计的三级模型及其基本内容。

概念模型设计是在较高的抽象层次上的设计， 其主要内容包括： 界定系统边界和确定主要的主题域。

逻辑模型设计的主要内容包括：分析主题域、确定粒度层次划分、确定数据分割策略、定义关系模式、定义记录系统。

物理数据模型设计的主要内容包括： 确定数据存储结构、 确定数据存放位置、确定存储分配以及确定索引策略等。在物理数据模型设计时主要考虑的因素有 ：I/O 存取时间、空间利用率和维护代价等。

提高性能的主要措施有划分粒度、数据分割、合并表、建立数据序列、引入冗余、生成导出数据、建立广义索引等。

7. 在数据挖掘之前为什么要对原始数据进行预处理？

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。

为数据挖掘算法提供完整、干净、准确、有针对性的数据，减少算法的计算量，提高挖掘效率和准确程度。

8. 简述数据预处理方法和内容。

① 数据清洗：包括填充空缺值，识别孤立点，去掉噪声和无关数据。

② 数据集成：将多个数据源中的数据结合起来存放在一个一致的数据存储中。需要注意不同数据源的数据匹配问题、数值冲突问题和冗余问题等。

③ 数据变换：将原始数据转换成为适合数据挖掘的形式。包括对数据的汇总、聚集、概化、规范化，还可能需要进行属性的重构。

④ 数据归约：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。

9. 简述数据清理的基本内容。

① 尽可能赋予属性名和属性值明确的含义；

② 统一多数据源的属性值编码；

③ 去除无用的惟一属性或键值 (如自动增长的 id) ；

④ 去除重复属性 (在某些分析中，年龄和出生日期可能就是重复的属性，但在某些时候它们可能又是同时需要的)

⑤ 去除可忽略字段 (大部分为空值的属性一般是没有什么价值的，如果不去除可能造成错误的数据挖掘结果)

⑥ 合理选择关联字段 (对于多个关联性较强的属性，重复无益，只需选择其中的部分用于数据挖掘即可，如价格、数据、金额)

⑦ 去掉数据中的噪音、填充空值、丢失值和处理不一致数据。

10. 简述处理空缺值的方法。

① 忽略该记录；

② 去掉属性；

③ 手工填写空缺值；

④ 使用默认值；

⑤ 使用属性平均值；

⑥ 使用同类样本平均值；

⑦ 预测最可能的值。

11. 常见的分箱方法有哪些？数据平滑处理的方法有哪些？

分箱的方法主要有：

① 统一权重法 (又称等深分箱法)

② 统一区间法 (又称等宽分箱法)

③ 最小熵法

④ 自定义区间法

数据平滑的方法主要有：平均值法、边界值法和中值法。

12. 何谓数据规范化？规范化的方法有哪些？写出对应的变换公式。

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 0.0～1.0），称为规范化。规范化的常用方法有：

(1) 最大－最小规范化：$x = \frac{max - min}{(max_0 - min_0)}(x_0 - min_0) + min$

(2) 零－均值规范化：$x = \frac{x_0 - \bar{x}}{\sigma_x}$

(3) 小数定标规范化：$x = x_0/10^{\alpha}$

13. 数据归约的方法有哪些？为什么要进行维归约？

① 数据立方体聚集

② 维归约

③ 数据压缩

④ 数值压缩

⑤ 离散化和概念分层

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。

14. 何谓聚类？它与分类有什么异同？

聚类是将物理或抽象对象的集合分组成为多个类或簇 (cluster) 的过程，使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。

聚类与分类不同，聚类要划分的类是未知的，分类则可按已知规则进行；聚类是一种无指导学习，它不依赖预先定义的类和带类标号的训练实例，属于观察式学习，分类则属于有指导的学习，是示例式学习。

15. 举例说明聚类分析的典型应用。

①商业：帮助市场分析人员从客户基本库中发现不同的客户群，并且用不同的购买模式描述不同客户群的特征。

②生物学：推导植物或动物的分类，对基于进行分类，获得对种群中固有结构的认识。

③WEB文档分类

④其他：如地球观测数据库中相似地区的确定；各类保险投保人的分组；一个城市中不同类型、价值、地理位置房子的分组等。

⑤聚类分析还可作为其他数据挖掘算法的预处理：即先进行聚类，然后再进行分类等其他的数据挖掘。聚类分析是一种数据简化技术，它把基于相似数据特征的变量或个案组合在一起。

16. 聚类分析中常见的数据类型有哪些？何谓相异度矩阵？它有什么特点？

常见数据类型有区间标度变量、比例标度型变量、二元变量、标称型、序数型以及混合类型等。 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。其特点是 d(i,j)=d(j,i) , d(i,i)=0 , d(j,j)=0 。如下所示：

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

17. 分类知识的发现方法主要有哪些？分类过程通常包括哪两个步骤？

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。分类的过程包括 2 步：首先在已知训练数据集上，根据属性特征，为每一种类别找到一个合理的描述或模型，即分类规则；然后根据规则对新数据进行分类。

18. 什么是决策树？如何用决策树进行分类？

决策树是用样本的属性作为结点，用属性的取值作为分支的树结构。它是利用信息论原理对大量样本的属性进行分析和归纳而产生的。决策树的根结点是所有样本中信息量最大的属性。树的中间结点是以该结点为根的子树所包含的样本子集中信息量最大的属性。决策树的叶结点是样本的类别值。

决策树用于对新样本的分类，即通过决策树对新样本属性值的测试，从树的根结点开始， 按照样本属性的取值， 逐渐沿着决策树向下， 直到树的叶结点，该叶结点表示的类别就是新样本的类别。决策树方法是数据挖掘中非常有效的分类方法。

19. 简述 ID3 算法的基本思想及其主算法的基本步骤。

首先找出最有判别力的因素，然后把数据分成多个子集，每个子集又选择最有判别力的因素进一步划分，一直进行到所有子集仅包含同一类型的数据为止。最后得到一棵决策树，可以用它来对新的样例进行分类。

主算法包括如下几步：

①从训练集中随机选择一个既含正例又含反例的子集 (称为窗口) ；

②用 “ 建树算法 ” 对当前窗口形成一棵决策树 ；

③对训练集 (窗口除外) 中例子用所得决策树进行类别判定，找出错判的例子 ；

④若存在错判的例子，把它们插入窗口，重复步骤②，否则结束。

20. 简述 ID3 算法的基本思想及其建树算法的基本步骤。

首先找出最有判别力的因素，然后把数据分成多个子集，每个子集又选择最有判别力的因素进一步划分，一直进行到所有子集仅包含同一类型的数据为止。最后得到一棵决策树，可以用它来对新的样例进行分类。

建树算法的具体步骤如下：

①对当前例子集合，计算各特征的互信息；

②选择互信息最大的特征 A_k；

③把在 A_k 处取值相同的例子归于同一子集， A_k 取几个值就得几个子集；

④对既含正例又含反例的子集，递归调用建树算法；

⑤若子集仅含正例或反例，对应分枝标上 P 或 N，返回调用处。

21. 设某事务项集构成如下表，填空完成其中支持度和置信度的计算。

事务 ID	项集	L2	支持度 %	规则	置信度 %
T1	A, D	A, B	33.3	A→B	50
T2	D, E	A, C	33.3	C→A	60
T3	A, C, E	A, D	44.4	A→D	66.7
T4	A, B, D, E	B, D	33.3	B→D	75
T5	A, B, C	C, D	33.3	C→D	60
T6	A, B, D	D, E	33.3	D→E	43
T7	A, C, D	,		,	
T8	C, D, E				
T9	B, C, D				

22. 从信息处理角度看，神经元具有哪些基本特征？写出描述神经元状态的 M-P方程并说明其含义。

基本特征：①多输入、单输出；②突触兼有兴奋和抑制两种性能；③可时间加权和空间加权；④可产生脉冲；⑤脉冲可进行传递；⑥非线性，有阈值。

M-P方程：$S_i = f(\sum_j W_{ij} S_j - \theta_j)$，$W_{ij}$ 是神经元之间的连接强度，θ_j 是阈值，$f(x)$ 是阶梯函数。

23. 遗传算法与传统寻优算法相比有什么特点？

① 遗传算法为群体搜索，有利于寻找到全局最优解；

② 遗传算法采用高效有方向的随机搜索，搜索效率高；

③ 遗传算法处理的对象是个体而不是参变量，具有广泛的应用领域；

④ 遗传算法使用适应值信息评估个体， 不需要导数或其他辅助信息， 运算速度快，适应性好；

⑤ 遗传算法具有隐含并行性，具有更高的运行效率。

24. 写出非对称二元变量相异度计算公式 (即 jaccard 系数)，并计算下表中各对象间的相异度。

测试项目 / 对 象	test-1	test-2	test-3	test-4	test-5	test-6
OBJ1	Y	N	P	N	N	N

OBJ2	Y	N	P	N	P	N
OBJ3	N	Y	N	Y	N	N
,	,	,	,	,	,	,

25. 简述 K-平均算法的输入、输出及聚类过程 (流程)。

输入：簇的数目 k 和包含 n 个对象的数据集。

输出：k 个簇，使平方误差准则最小。

步骤：

① 任意选择 k 个对象作为初始的簇中心；

② 计算其它对象与这 k 个中心的距离，然后把每个对象归入离它“最近”的簇；

③ 计算各簇中对象的平均值，然后重新选择簇中心 (离平均值“最近”的对象值)；

④ 重复第 2 第 3 步直到簇中心不再变化为止。

26. 简述 K-中心点算法的输入、输出及聚类过程 (流程)。

输入：结果簇的数目 k，包含 n 个对象的数据集

输出：k 个簇，使得所有对象与其最近中心点的相异度总和最小。

流程：

① 随机选择 k 个对象作为初始中心点；

② 计算其它对象与这 k 个中心的距离，然后把每个对象归入离它“最近”的簇；

③ 随机地选择一个非中心点对象 Orandom，并计算用 Orandom 代替 Oj 的总代价 S；

④ 如果 S<0，则用 Orandom 代替 Oj，形成新的 k 个中心点集合；

⑤ 重复迭代第 3、4 步，直到中心点不变为止。

27. 何谓文本挖掘？它与信息检索有什么关系 (异同)。

文本挖掘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。它与信息检索之间有以下几方面的区别：

① 方法论不同：信息检索是目标驱动的，用户需要明确提出查询要求；而文本挖掘结果独立于用户的信息需求，是用户无法预知的。

② 着眼点不同：信息检索着重于文档中字、词和链接；而文本挖掘在于理解文本的内容和结构。

③ 目的不同：信息检索的目的在于帮助用户发现资源，即从大量的文本中找到满足其查询请求的文本子集；而文本挖掘是为了揭示文本中隐含的知识。

④ 评价方法不同：信息检索用查准率和查全率来评价其性能。而文本挖掘

采用收益、置信度、简洁性等来衡量所发现知识的有效性、可用性和可理解性。

⑤ 使用场合不同： 文本挖掘是比信息检索更高层次的技术， 可用于信息检索技术不能解决的许多场合。一方面，这两种技术各有所长，有各自适用的场合；另一方面，可以利用文本挖掘的研究成果来提高信息检索的精度和效率， 改善检索结果的组织， 使信息检索系统发展到一个新的水平。