**数据挖掘的其中一套题**

**一·填空题(5小题,每空1分,共15分)**

1.关联规则的经典算法包括（ ）和（ ）

其中（ ）的效率更高.

2.分类器设计阶段包含三个过程,（ ）（）和（ ）

3.分类问题中常用的评价准则有（  ）（）和（ ）

4.支持向量机常用的核函数有（）（ ）和（ ）

5.连续型属性的数据样本之间的距离有（ ）（  ）和（ ）

6.层次聚类方法包括（）和（）两种层次聚类方法

7.聚类分析包括（  ）（ ）（）和（ ）四种类型描述属性的相似度计算方法。

8.划分聚类方法对数据集进行聚类时包含三个要点:

9.数据仓库是一个（）（）（），()数据集合。

10.数据处理通常分为两大类：（）和（）

11.使用星形模式可以从一定程度上（）查询效率，因为星型模式中数据的组织已经经过（） 主要数据都在庞大的（）中。

**二、简述题(5小题,每题5分,共25分)**

1.请简述数据挖掘中关联规则 Apriori算法的思想。

2.请比较PCA和LDA的区别。

1. 请分析特征选择和特征提取有何区别?

4.聚类和分类有什么区别和联系?

5.TF-IDF算法是什么,有什么实际意义?

1. 数据挖掘与统计的区别与联系?

7.什么是关联规则? 关联规则的应用有哪些?

关联规则（Association Rules）是反映一个事物与其他事物之间的相互依存性和关联性

1. 什么是分类?分类的应用领域有哪些?

9.什么是聚类分析?聚类分析的应用领域有哪些?

10.什么是数据仓库的数据ETL过程?

答：数据的ETL过程就是负责将操作型数据转换成调和数据的过程，包括抽取、清洗、加载与索引。这些步骤可以进行不同的组合，如，可以将数据抽取与清洗组合为一个过程，或者将清洗和转换组合在一起。

11.什么是信息包图法?它为什么适用于数据仓库的概率模型设计？

答：①信息包图法就是在一张平面表格上描述元素的多维性，其中每一个维度用平面表格的一列表示，而细化本列的对象就是类别，平面表格的最后一行即为指标度量表。

②因为信息包图法是一种自上而下的数据建模方法，即从用户的观点开始设计，站在管理者的角度把焦点集中在企业的一个或几个主题上，着重分析主题所涉及数据的多维特性，这种自上而下的方法几乎考虑了所有的信息源，以及这些信息源影响业务活动的方式。

12.关联规则的分类有哪些?关联规则挖掘的步骤包括什么？

答：分类：①基于规则中涉及到的数据的维数，关联规则可以分为单维和多维的。

②基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

③基于规则中处理的变量的类型不同，关联规则可以分为布尔型和数值型。

步骤：①找出交易数据库中所有大于或等于用户指定的最小支持度的频繁项集；

②利用频繁项集生成所需要的关联规则，更具用户设定的最小可信度进行取舍，产生强关联规则。

13.请描述k-means聚类算法的操作步骤

　　(1)从数据中选择k个对象作为初始聚类中心;

　　(2)计算每个聚类对象到聚类中心的距离并划分类别；

(3)计算划分的每个类别的均值作为新的聚类中心

(4)计算每个聚类对象到新的聚类中心的距离并划分到距离最近的类别中；

(5)重复34直到各个聚类不再发生变化

14.贝叶斯网路的三个主要议题是什么?

答：①预测；②诊断；③历史数据训练

15.怎样从历史数据中训练出结点之间的条件概率或联合条件概率?

答：① 要训练条件概率P（B|A），可以在历史数据中统计A发生的次数T（A），然后统计在A发生的数据中B发生的次数T（A，B），条件概率P（B|A）=T(B)/T(A).

②要训练联合条件概率P（C|A，B），可以在历史数据中统计A,B共同发生的次数

T（A,B）,然后再A,B共同发生的数据中统计C发生的次数T（A,B,C），联合条件概率

P（C|A，B）=T（A,B,C）/T（C,B）

16.前馈网络和递归网络有什么区别?

答：前馈网络和递归网络的本质区别是网络的某些输出是不是循环作为网络的输入。前馈网络的所有输出都不能作为输入，而递归网络的某些循环可以作为网络的输入。

17.简答遗传算法包括哪些步骤？

答：遗传算法步骤：

SGA的基本流程如下：

（1）初始化，产生初始种群。

（2）个体评价，即计算种群中每个个体的适应度。

（3）按选择概率Ps，执行选择算子，从当前种群中选择部分个体进入下一代种群。

（4）按交叉概率Pc，执行交叉算子。

（5）按变异概率Pm，执行变异算子。

（6）若满足设定的终止条件，则输出种群中适应度最优的个体作为问题的最优解或满意解，否则执行（2）

**三、计算题(3小题,每题15分,共45分)**

**（书上的）**

**四、论述题(1小题,共15分)**

列举几项你知道的数据挖掘应用,并论述数据挖掘在其中的作用?

10.简述你对数据仓库未来发展趋势的看法。

数据仓库技术的发展包括数据抽取、存储管理、数据表现和方法论等方面。

在数据抽取方面，未来的技术发展将集中在系统集成化方面。进行标准化的统一管理，以适应数据仓库本身或数据源可能的变化，使系统更便于管理和维护。

在数据管理方面，未来数据库厂商明确推出数据仓库引擎，作为数据仓库服务器产品与数据库服务器并驾齐驱。在这一方面，带有决策支持扩展的并行关系数据库将最具发展潜力。

在数据表现方面，数理统计的算法和功能将普遍集成到联机分析产品中，并与Internet/Web技术紧密结合。按行业应用特征细化的数据仓库用户前端软件将成为产品作为数据仓库解决方案的一部分。

数据仓库实现过程的方法论将更加普及，将成为数据库设计的一个明确分支，成为管理信息系统设计的必备。

11.请列出3中数据仓库产品，并说明其优缺点?

答：（1）IBM提供了基于可视化数据仓库的商业智能解决方案，其中，VW是一个功能很强的集成环境，既可用于数据仓库建模和源数据管理，又可用于数据抽取、转换、装载和调度。

1. Oracle数据仓库解决方案主要包括Oracle Express和Oracle Discoverer两个部分。Oracle Express利用多维模型，存储和管理多维数据库或多维高速缓存，同时也能够访问多种关系数据库。Oracle Discoverer即查询工具是专门为最终用户设计的，使用方便简洁分为最终用户版和管理员版。

（3）Microsoft将OLAP功能集成到SQL Server数据库中，其解决方案包括BI平台、BI终端工具、BI门户和BI应用四个部分。BI平台是BI解决方案的基础。