**《数据仓库与数据挖掘》作业3**

1. **单项选择题**

1. 数据规范化的目的是（ ）

A、去掉数据中的噪声

B、对数据进行汇总和聚集

C、使用概念分层，用高层次概念替换低层次“原始”数据

**D、将属性按比例缩放，使之落入一个小的特定区间**

2. 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

3. 数据挖掘应用和一些常见的数据统计分析系统的最主要区别在于：（ ）。

A、所涉及的算法的复杂性； **B、所涉及的数据量；**

C、计算结果的表现形式； D、是否使用了人工智能技术

4. 密度聚类方法可以用于过滤噪声孤立点数据，发现（ ）而不仅仅是球形簇。

**A、任意形状** B、矩形 C、网格形状 D、椭圆形状

5. （ ）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

6. DBSCAN在最坏情况下的时间复杂度是（ ）。

A、O(m) **B、O(m2)**  C、O(log m) D、O(m\*log m)

7. 孤立点挖掘适用于下列哪种场合（ ）。

A、目标市场分析 B、购物篮分析 C、模式识别  **D、信用卡欺诈检测**

8. 噪声数据的处理方法不包含（ ）方法。

A、分箱 B、回归 C、聚类  **D、关联规则**

9. 关于K均值和DBSCAN的比较，以下说法不正确的是（ ）。

**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。

10. 要建立预测模型，需要准备建模数据集，即训练集，以下四条描述训练集正确的是（ ）。

A、数据越多越好 B、测试集是训练集数据的一部分

**C、尽可能多的适合数据**  D、以上三条都正确

1. **计算题**

1. 请利用Apriori算法求项目集I={ I1，I2，I3，I4，I5}中，根据下列事务表：

|  |  |
| --- | --- |
| TID | 项ID的列表 |
| T100  T200  T300  T400  T500  T600  T700  T800  T900 | I1，I2，I4  I1，I3  I2，I4  I1，I4，I5  I1，I3  I2，I3  I1，I4  I1，I2，I4，I5  I1，I2，I3 |

完成：

1）假定最小支持度设定为2，请找出了所有的最大频繁项集；

2）假定最小置信度设定为65％，请求出最大项集的关联规则；

3）分析所得到的结论。

2. 采用K-means方法对给定的如下5个点进行聚类，（0, 2）（0, 0）（1.5, 0）（5, 0）（5, 2），要求簇的数量k=2，且初始簇心为（0, 2）（0, 0）算法执行后两个簇的结果是？

1. **分析题**

1. 网络安全已经成为电子商务安全的基础和重要组成部分, 因此必须找到保护系统的最好办法。网络入侵定义为: 任何试图破坏资源的完整性、机密性和可用性的行为。基于机器学习设计一种入侵检测的方法。

1）请设计出该策略的具体步骤（绘出流程图）

2）请写出该方法用到的算法的输入、输出和过程

3）该算法还适用于什么场合？