中国矿业大学银川学院期末考试试题

2010 至 2011 学年第 2 学期

考试科目 数据仓库与数据挖掘 学分 2 年级 2008

系 机电动力与信息工程系 专业 计算机

一、 填空题（15 分）

1.数据仓库的特点分别是 面向主题 、 集成 、 相对稳定 、反映历史变化。

2.元数据是描述数据仓库内数据的结构和建立方法的数据。根据元数据用途的不同可将元数据分为 技术 元数据和 业务 元数据两类。

3.OLAP 技术多维分析过程中，多维分析操作包括 切片 、 切块 、 钻取 、 旋转 等。

4.基于依赖型数据集市和操作型数据存储的数据仓库体系结构常常被称为“ 中心和辐射 ”架构，其中 企业级数据仓库 是中心，源数据系统和数据集市在输入和输出范围的两端。

5.ODS 实际上是一个集成的、 面向主题的 、 可更新的 、 当前值的 、 企业级的 、详细的数据库，也叫运营数据存储。

二、 多项选择题（10 分）

6.在数据挖掘的分析方法中，直接数据挖掘包括（ A C D ）

A 分类 B 关联 C 估值 D 预言

7.数据仓库的数据 ETL 过程中，ETL 软件的主要功能包括（ A B C ）

A 数据抽取　　B 数据转换　　C 数据加载　　D 数据稽核

8.数据分类的评价准则包括（　ABCD　）

A 精确度　　B 查全率和查准率　　C F-Measure　　D 几何均值

9.层次聚类方法包括（　　BC　　）

A 划分聚类方法　B 凝聚型层次聚类方法　C 分解型层次聚类方法　D 基于密度聚类方法

10.贝叶斯网络由两部分组成，分别是（　A D　）

A 网络结构　　B 先验概率　　C 后验概率　　D 条件概率表

三、　计算题（30 分）

11.一个食品连锁店每周的事务记录如下表所示，其中每一条事务表示在一项收款机业务中卖出的项目，假定 sup_{min}=40%，$conf_{min}$=40%，使用 Apriori 算法计算生成的关联规则，标明每趟数据库扫描时的候选集和大项目集。（15 分）

事务	项目	事务	项目
T1	面包、果冻、花生酱	T4	啤酒、面包
T2	面包、花生酱	T5	啤酒、牛奶
T3	面包、牛奶、花生酱		

解：（1）由 I={面包、果冻、花生酱、牛奶、啤酒}的所有项目直接产

生 1-候选 C_1，计算其支持度，取出支持度小于 sup_{min} 的项集，形成 1-频繁集 L_1，如下表所示：

项集 C_1	支持度	项集 L_1	支持度
{面包}	4/5	{面包}	4/5
{花生酱}	3/5	{花生酱}	3/5
{牛奶}	2/5	{牛奶}	2/5
{啤酒}	2/5	{啤酒}	2/5

(2)组合连接 L_1 中的各项目，产生 2-候选集 C_2，计算其支持度，取出支持度小于 sup_{min} 的项集，形成 2-频繁集 L_2，如下表所示：

项集 C_2	支持度	项集 L_2	支持度
{面包、花生酱}	3/5	{面包、花生酱}	3/5

至此，所有频繁集都被找到，算法结束，

所以，confidence（{面包}→{花生酱}）=（4/5）/（3/5）=4/3> $conf_{min}$

confidence（{ 花生酱}→{面包}）=（3/5）/（4/5）=3/4> $conf_{min}$

所以，关联规则{面包}→{花生酱}、{ 花生酱}→{面包}均是强关联规则。

12.给定以下数据集（2 ，4，10，12，15，3，21），进行 K-Means 聚类，设定聚类数为 2 个，相似度按照欧式距离计算。（15 分）

解：（1）从数据集 X 中随机地选择 k 个数据样本作为聚类的出示代表点，每一个代表点表示一个类别，由题可知 k=2，则可设 m_1=2，m_2=4：

（2）对于 X 中的任意数据样本 x_m（1<x_m<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m_1=2 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为 2，8，10，13，1，19。

当 m_2=4 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-2，6，8，11，-1，17。

最小距离是 1 或者-1 将该元素放入 m_1=2 的聚类中，则该聚类为（2，3），另一个聚类 m_2=4 为（4，10，12，15，21）。

（3）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到 k 个均值代表点：m_1=2.5，m_2=12：

（4）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m_1=2.5 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-0.5，0.5，1.5，7.5，9.5，12.5，18.5。

当 m_2=12 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-10，-9，-8，2，3，9。

最小距离是 1.5 将该元素放入 m_1=2.5 的聚类中，则该聚类为（2，3，4），另一个聚类 m_2=12 为（10，12，15，21）。

（5）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到 k 个均值代表点：m_1=3， m_2=14.5：

（6）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中：当 m_1=3 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-1，1，7，9，12，18，。

当 m_2=14.5 时，样本（2 ，4，10，12，15，3，21）距离该代表点的距离分别为-12.58，-11.5，-10.5，-4.5，-2.5，0.5，6.5。

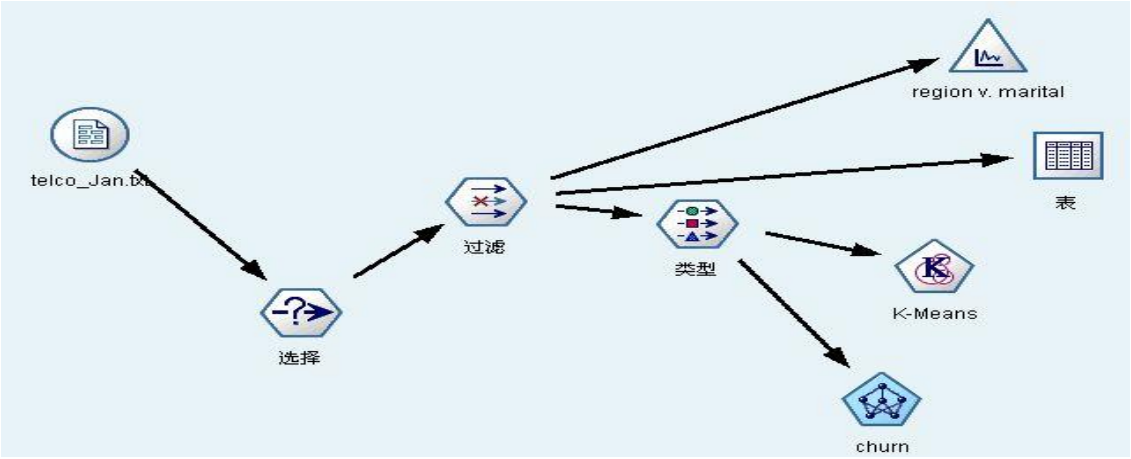
最小距离是 0.5 将该元素放入 m_1=3 的聚类中，则该聚类为（2，3，4），另一个聚类 m_2=14.5 为（10，12，15，21）。

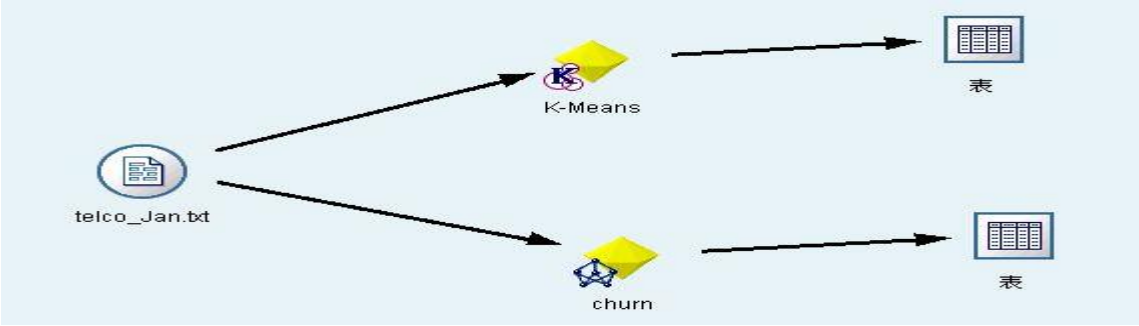
至此，各个聚类不再发生变化为止，即误差平方和准则函数的值达到最优。

四．设计题（45 分）

13.按照题目给定的 3 个数据文件，任选一个建立数据流图，要求至少包括记录选项、字段选项、图形结点各一个。任选关联规则 Apriori 算法、贝叶斯网络、K-Means 聚类、决策树 C5.0（C4.5）算法、神经

网络中的一个进行挖掘，并给出数据流图。（10 分）





14.对以上数据流图中使用的每个结点做一简短说明。（10 分）

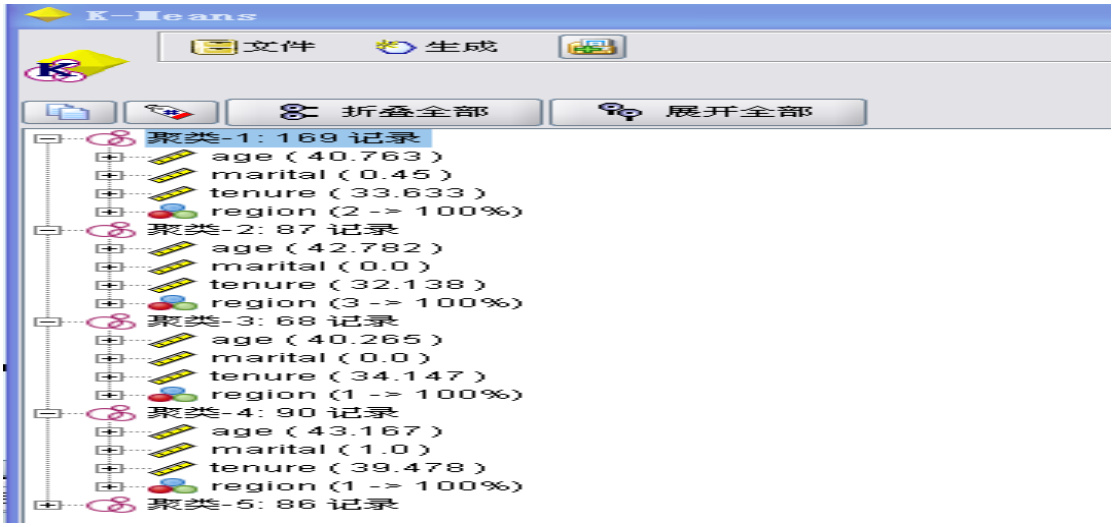
选择：age>25.

过滤：过滤后的字段。Region,tenure,age,marital,churn.

类　　型　　：

字段	类型	值	缺失	检查	方向
region	集	1,2,3		无	输入
tenure	范围	[1,72]		无	输入
age	范围	[18,77]		无	输入
marital	范围	[0,1]		无	输入
churn	范围	[0,1]		无	输出

15.给出以上数据流图中模型的执行结果（生成模型完全展开后的数据），对于执行结果太多的，可节选部分结果。（10 分）



16.对以上模型生成的结果做一简要的分析，包括算法采用的基本原理、数学模型、算法步骤等。(15 分)

答：k-means 聚类算法基本原理：将各个聚类子集内的所有数据样本的均值作为该聚类的代表点，算法的主要思想是通过迭代过程把数据划分为不同的类别，使得评价聚集类性能的准则函数达到最优，从而使生成的每个聚集类的紧凑，类间独立。

操作步骤：

输入：数据集,其中的数据样本只包含描述属性，不包含类别属性。聚类个数 K

输出：

（1）从数据集 X 中随机地选择 k 个数据样本作为聚类的出示代表点，每一个代表点表示一个类别

（2）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类

别中

（3）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到 k 个均值代表点

（4）对于 X 中的任意数据样本 xm（1<xm<total），计算它与 k 个初始代表点的距离，并且将它划分到距离最近的初始代表点所表示的类别中

（5）重复 3.4，直到各个聚类不再发生变化为止。即误差平方和准则函数的值达到最优