作业三

姓名：陈庆辉

学号：1714080902201

班级：17软件工程2班

请简述数据仓库中属性的类型及相应的一些度量方法。

答：**标称属性**：1. 标称属性的值是一些符号或事物的名称；

2. 每个值代表某种类别、编码、状态，因此标称属性又被看做是分类的（categorical）；

3. 标称属性的值不具有有意义的序，而且不是定量的。（也就是说，给定一个对象集，找出这种属性的均值没有意义）。

**二元属性**：1. 二元属性是一种标称属性，只有两个状态：0 或 1，其中 0 通常表示该属性不出现，1 表示出现；

2. 二元属性又称布尔属性，如果两种状态对应的是 true 和 false。

**序数属性**：1.属性可能的值之间具有有意义的序；

2. 序数属性的中心趋势可以用它的众数和中位数表示；

3. 不适合使用平均值。

**数值属性**：1. 是定量的可度量的量，用整数或实数表示；

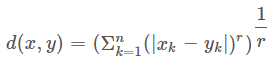
2. 可以是区间标度的或比率标度的。

**连续与离散属性**：1. 连续属性，通常在一个实数区间内取值，因此其取值个数理论上是不可数无限的；

2. 离散属性，是指该属性可以取有限或可数无限个不同的值，其取值可用字母或自然数表示，也可用单词或短语表示。

**度量方法**：

闵可夫斯基距离：



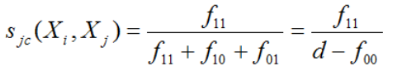
马氏距离：



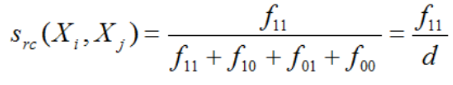
简单匹配系数：



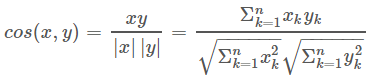
Jaccard系数：



Rao系数：



余弦距离：



作业四

姓名：陈庆辉

学号：1714080902201

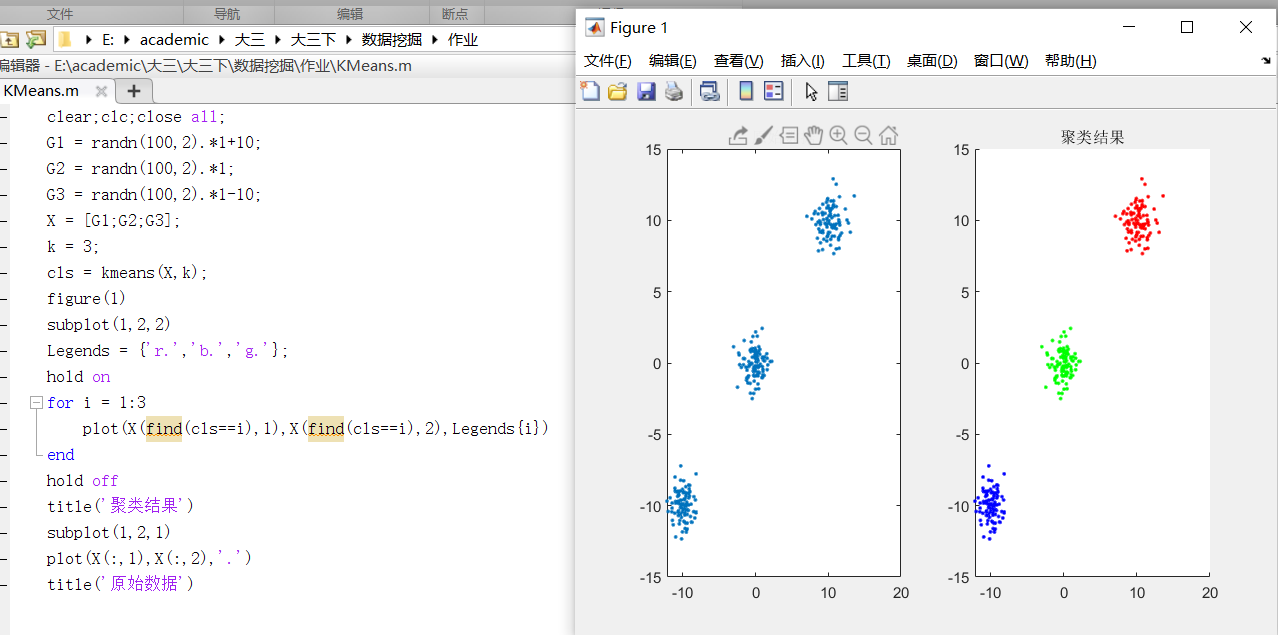
班级：17软件工程2班

请叙述和分析K-MEANS聚类算法并给出实例演示。

答：K-MEANS算法也称为K平均算法，它采用距离作为相异度的评价指标，以簇内差异函数W(C)作为聚类质量的优化目标函数，即将所有数据对象到它的簇中心的距离平方和作为目标函数，算法寻找最优聚类的策略是使目标函数达到最小值（簇中心不变化等价于W(C)达最小）。

实例：

生成观察点，假设K=3。第一轮随机产生3个簇的中心，然后分别计算每个观察点到这些中心的距离，将观察结果分配到最近的中心。然后使用这些组成员的平均位置来更新中心位置。第二轮重新计算……直到中心收敛到固定地点。



自我评价：

1. 写作是否工整，回答问题是否完整？

答：写作工整，回答问题完整。

2. 下一阶段学习目标？

答：理论课已上完，接下来重心放在实验课，用实际行为去巩固理论知识。