

**本科实验报告**

课程名称： 数据挖掘

课程编号： 08060116

学生姓名： 邝庆璇

学号： 2016051598

学院： 信息科学与技术学院

系： 计算机系

专业： 计算机科学与技术

指导教师： 刘波

教师单位： 暨南大学计算机系

开课时间：2018~ 2019学年度 第2期

**暨南大学教务处**

**2019年06月16日**

数据挖掘 **课程实验项目目录**

学生姓名：邝庆璇 学号：2016051598

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 实验项目编号 | 实验项目名称 | \*实验项目类型 | 成绩 | 指导教师 |
| 1 | **0806011601** | **关联分析** | **综合** |  | **刘波** |
| 2 | **0806011602** | **分类** | **综合** |  | **刘波** |
| 3 | **0806011603** | **聚类** | **综合** |  | **刘波** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

\*实验项目类型：演示性、验证性、验证性、设计性实验。

\*此表由学生按顺序填写。

**暨南大学本科实验报告专用纸**

课程名称 数据挖掘 成绩评定

实验名称 实验一 关联分析 指导教师 刘波

试验编号 0806011601 类型 综合性

学生姓名 邝庆璇 学号 2016051598

学院 信息科学技术学院 系 计算机 专业 计算机科学与技术

**1、实验目的：**

（1）理解关联分析原理

（2）理解置信度、支持度等概念

（3）掌握Apriori算法

1. **实验内容**

（0）对航空公司的乘客数据进行预处理（清洗、特征筛选），并得到5个关键的指标：客户关系时长L、消费时间间隔R、消费频率F、飞行里程M和折扣系数的平均值C，并进行归一化。

（1）用Apriori算法，对航空信息的这5个指标进行关联分析；

（2）输出并比较设定不同置信度下的实验结果。

1. **实验原理**

Apriori算法的思路如书上P94页所言，即按照一定的置信度和支持度，每次迭代挑选k频繁项集。每次生成k频繁项集的方法：把只差1项的两个k-1频繁项集连接在一起，检查连接后的k项集的每个候选子集是不是都是频繁的；若是，则把该k项集归为k频繁项集。

·两条依据：

1）若集合存在某一子集不是频繁项集，则该集合本身也不是频繁的。

2）若集合本身是频繁的，则其任意非空子集都是频繁的。

·算法如下：

输入：数据集合D，支持度阈值α

输出：最大的频繁k项集

1）扫描整个数据集，得到所有出现过的数据，作为候选频繁1项集。k=1，频繁0项集为空集。

2）挖掘频繁k项集

a) 扫描数据计算候选频繁k项集的支持度

b) 去除候选频繁k项集中支持度低于阈值的数据集,得到频繁k项集。如果得到的频繁k项集为空，则直接返回频繁k-1项集的集合作为算法结果，算法结束。如果得到的频繁k项集只有一项，则直接返回频繁k项集的集合作为算法结果，算法结束。

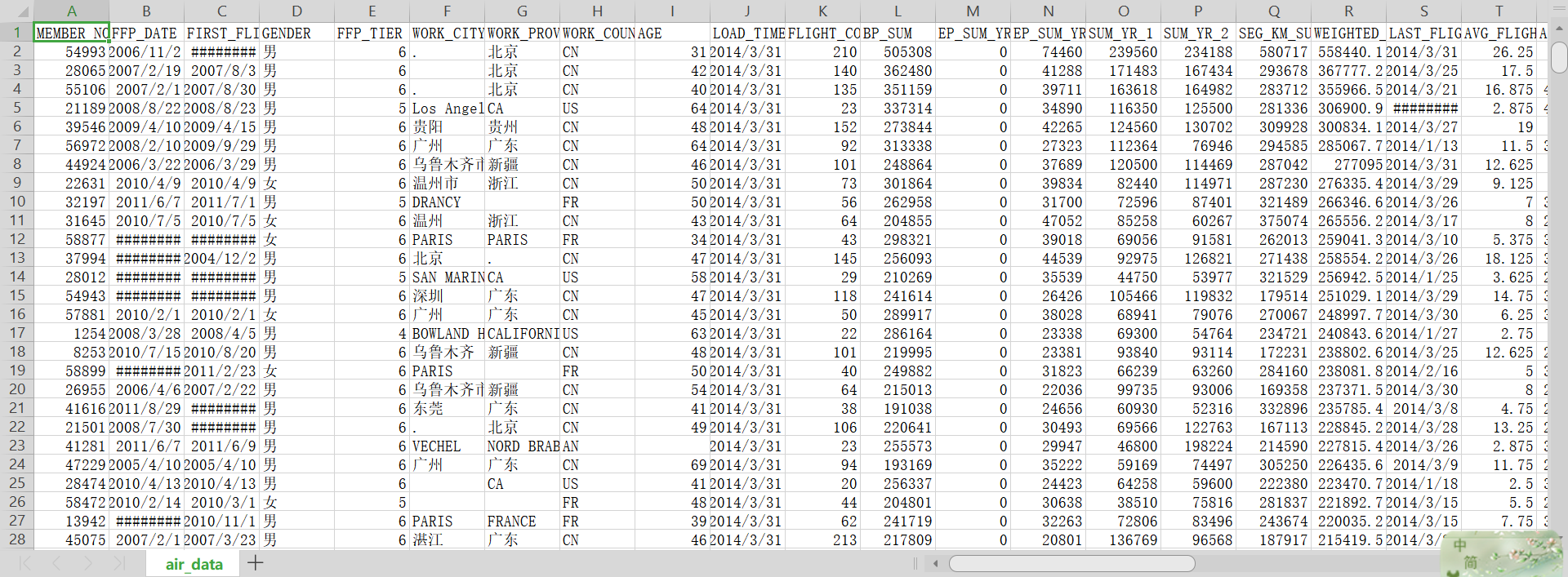
　 c) 基于频繁k项集，连接生成候选频繁k+1项集。

3） 令k=k+1，转入步骤2。

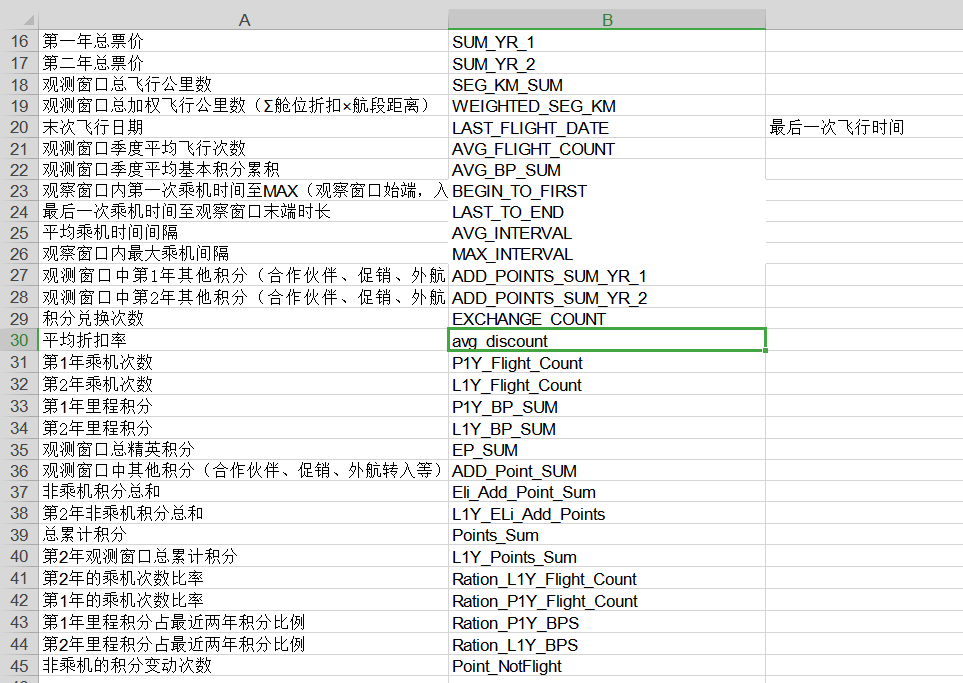
**3、实验过程：**

**步骤0**. 预处理数据集

航空旅客的信息的原始数据集及其含义如下：

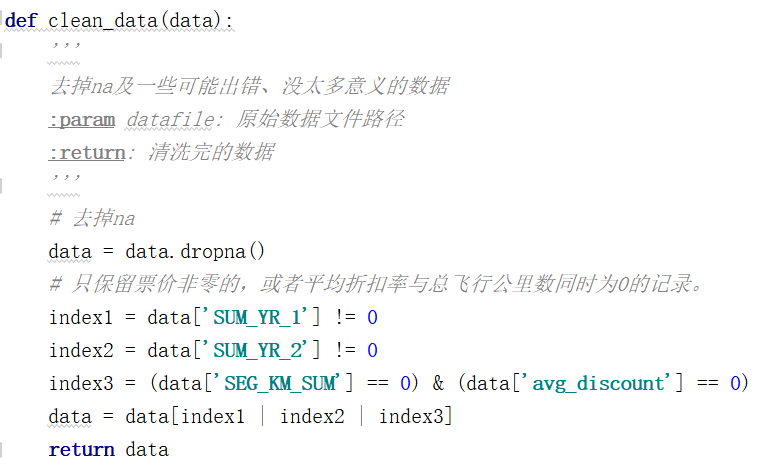


（数据集，由于列太多了，不太完全）

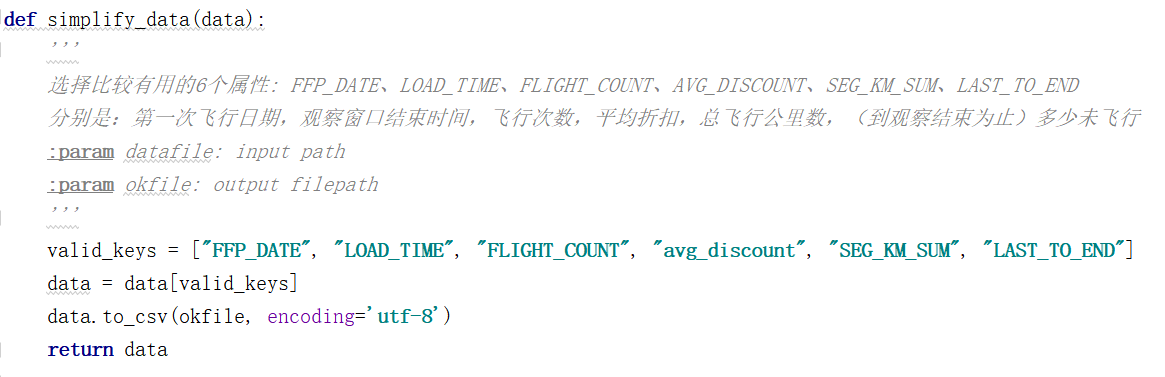


（意义）

0.1 清洗掉空的和无意义的行



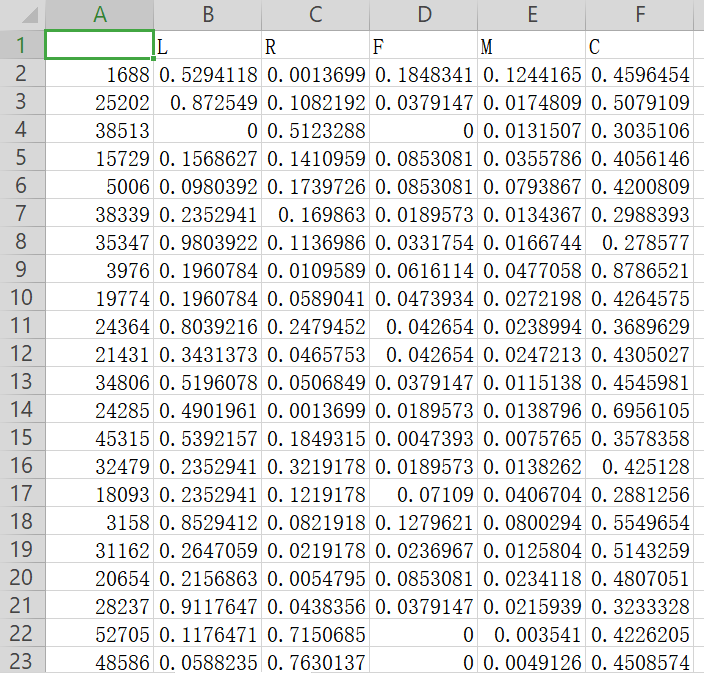
0.2 选取其中6个比较有代表性的特征



0.3 得到5个对客户价值的评价指标：客户关系时长L、消费时间间隔R、消费频率F、飞行里程M和折扣系数的平均值C，并归一化



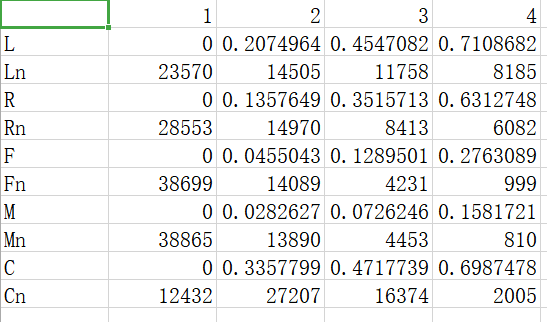
步骤0处理完成后，得到数据集如下：



（为了实验跟快速地进行，只随机地选择了原数据集5万多条数据中的部分保存。第一列是对应原数据集里面的编号。）

**步骤1** 把上述的LRFMC的每一个指标，都给离散化。

1.1 通过聚类，给每个指标都划成4个区间，返回该区间的最小值以及样本个数。如下：



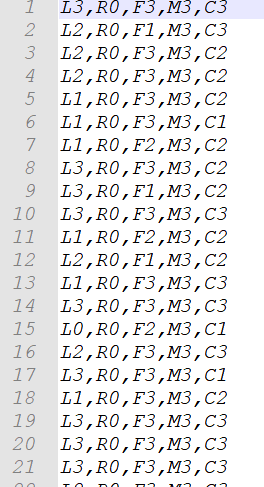
LRFMC每一行代表一个指标最小值，Xn每行代表X指标的样品个数

列代表每个指标划分成的4类。在rules.csv里。

代码如下：

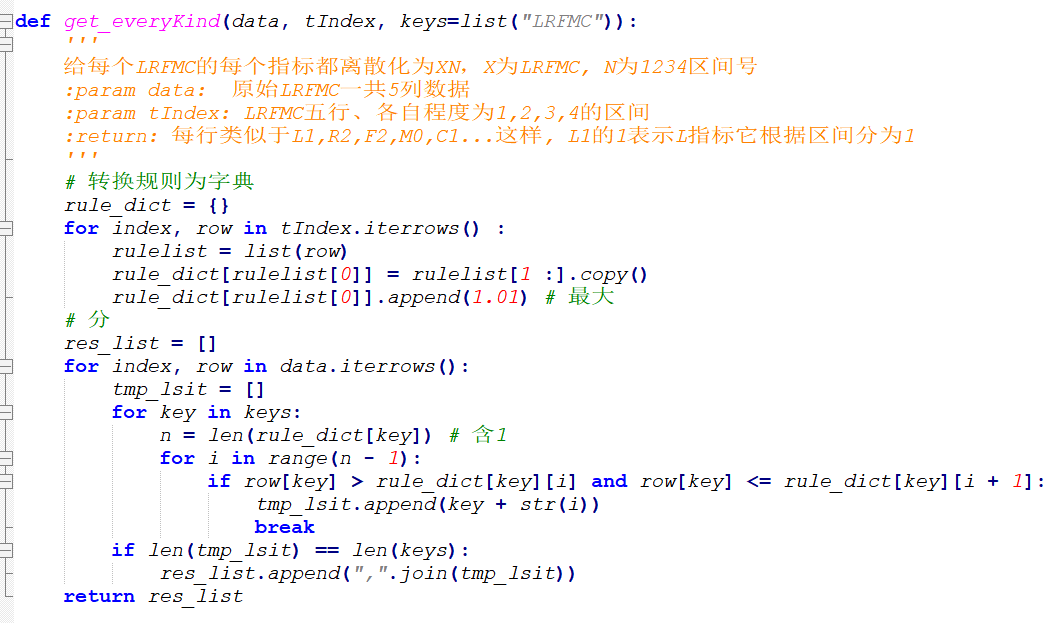


1.2 实现指标离散化为标号，如L1, L2, L3, L4, M1, M2,....

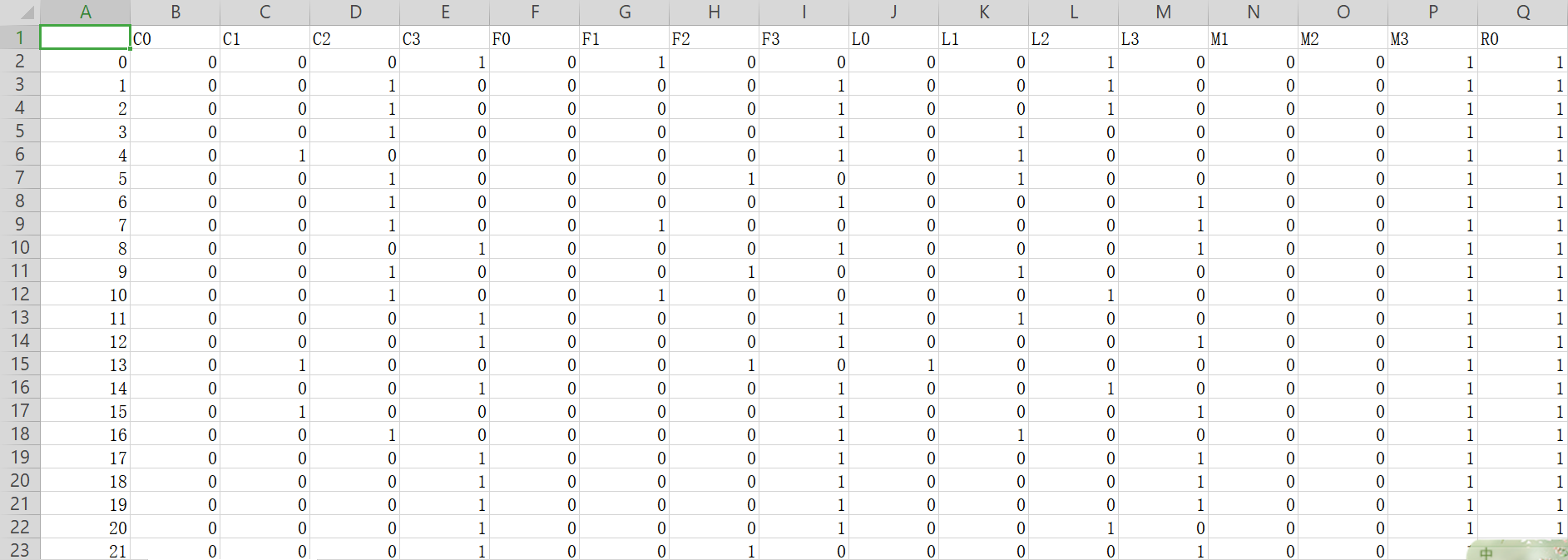


（dis.txt里）

代码如下：

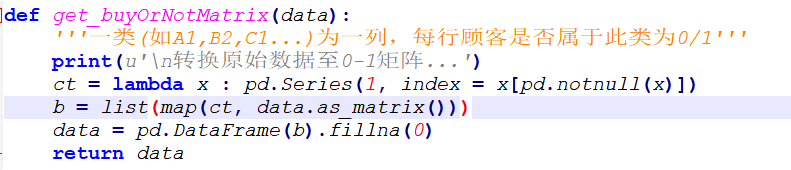


1.3 把数据转换为类似于one-hot矩阵



（association\_matrix.csv文件里，每一行代表每一行数据）

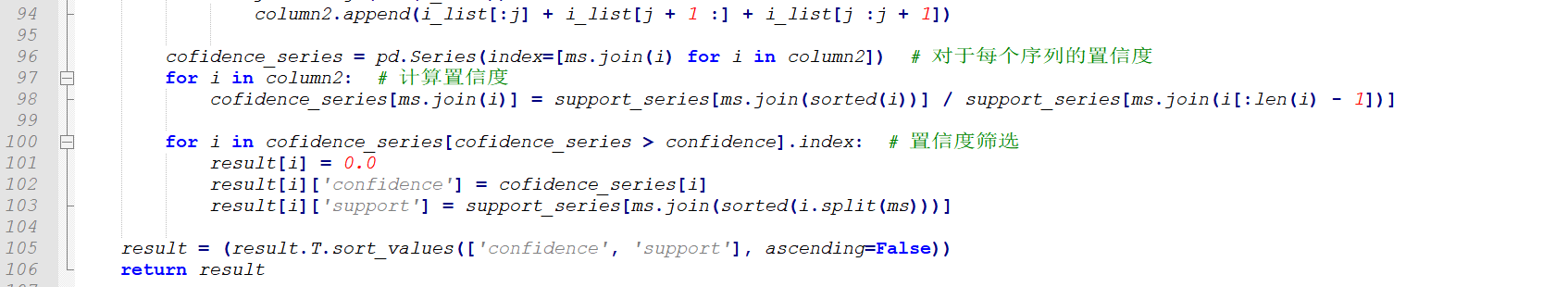
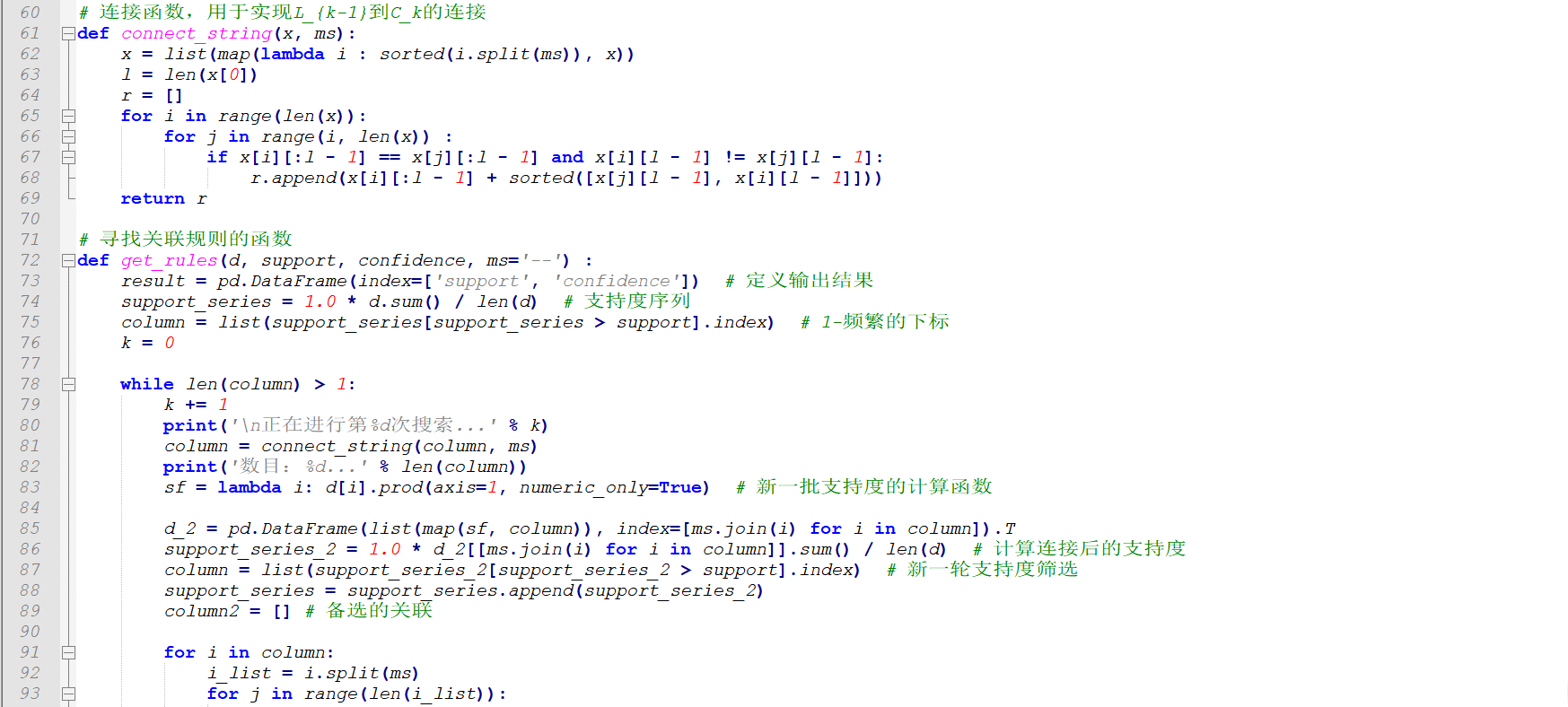
代码如下：



**步骤2**：进行关联分析

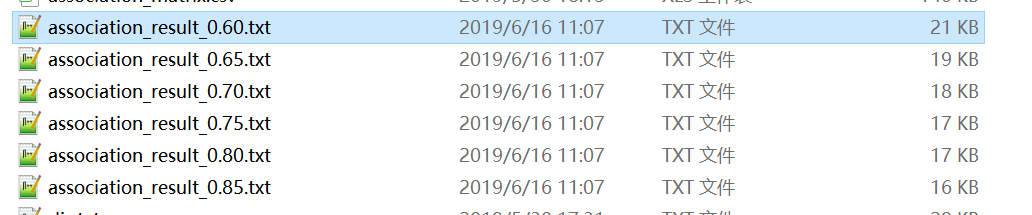
最低支持度为0.06，分别尝试最小置信度为0.60, 0.65, 0.70, 0.75, 0.80, 0.85下，寻找频繁项集。

代码如下：

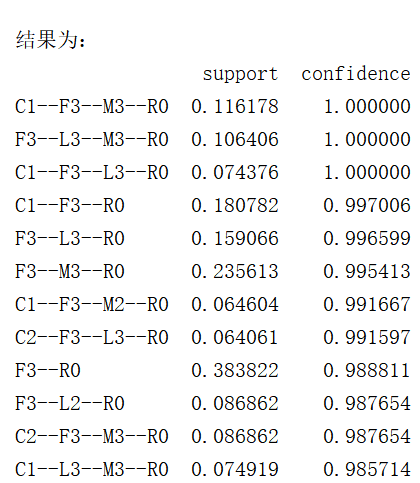


**5、实验结果：**

频繁项集结果保存在以下6个文件夹中：



以置信度最小为0.85的为例：



**6、实验总结：**

1. Apriori算法理解起来似乎不算难，而且也有书上的算法来参考但实现起来也有种种问题要克服。个人代码能力需要提供。

2.数据预处理和特征选择是一本数据挖掘相关的书提供的思路，我发现其实在真正关联分析前的这些步骤，也是很重要的，并且对于我这种对python里面pandas，numpy这些库不算熟的人来说，也要耗一点精力和时间。

3.收获：对关联规则算法更加理解加深了，并且对python里上述pandas, numpy, sklearn这些数据挖掘处理的常用库，更加熟练了。后面两个实验做起来比这个顺利了一点。

4. 一个注意的点：用pandas读csv和存csv时，一定要用.iloc[:, 1:]，把第0列index去掉！不然连索引index也会被当成数据，然后实验结果会异常！以后我写代码时也应注意这一点。

**暨南大学本科实验报告专用纸**

课程名称 数据挖掘 成绩评定

实验名称 实验二 分类 指导教师 刘波

试验编号 0806011602 类型 综合性

学生姓名 邝庆璇 学号 2016051598

学院 信息科学技术学院 系 计算机 专业 计算机科学与技术

**1、实验目的：**

（1）理解数据挖掘分类算法的原理

（2）掌握并熟悉常用分类算法

**2、实验内容**

利用K近邻算法，根据实验一中预处理得到的LRMC，把乘客分为“常客”和“并非常客”两类。

**3、实验原理**

如书上P128所言，基本步骤为：

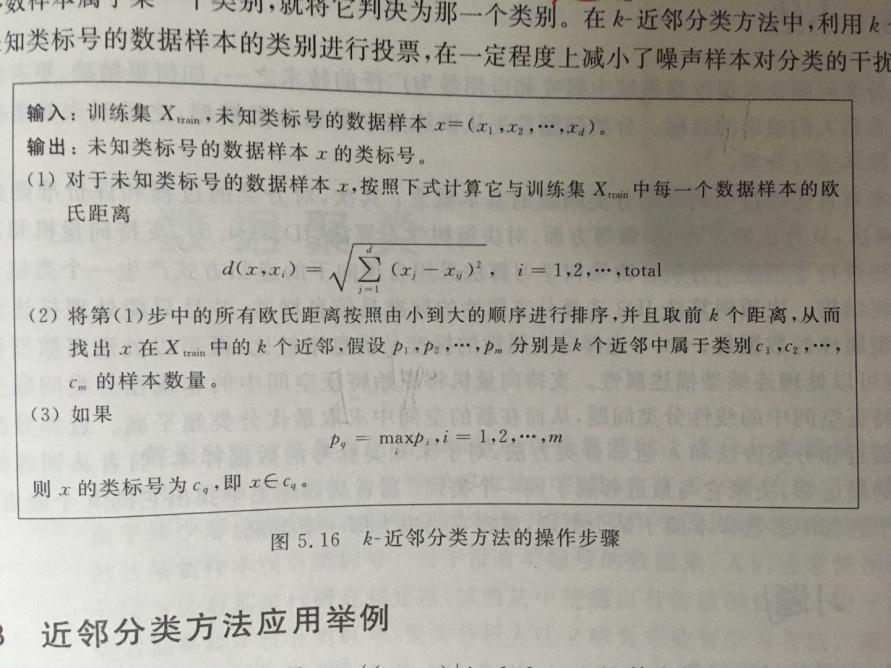
1）计算测试数据与各个训练数据之间的距离；

2）按照距离的递增关系进行排序；

3）选取距离最小的K个点；

4）确定前K个点所在类别的出现频率；

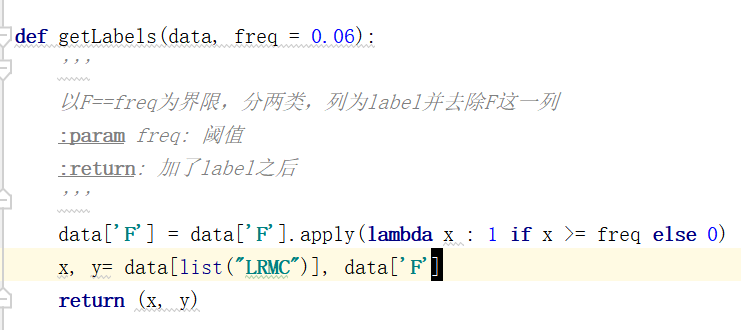
5）返回前K个点中出现频率最高的类别作为测试数据的预测分类。



（课本P129）

**4、实验过程：**

（1）把实验一里面预处理好的、含有每个顾客LFMRC这5个指标的数据（air\_LRFMC.csv），按照F（Frequency，顾客搭乘频率）是否大于0.06，把顾客分为“常客”和“并非常客”两类，并据此把F指标分别离散化为1和0。

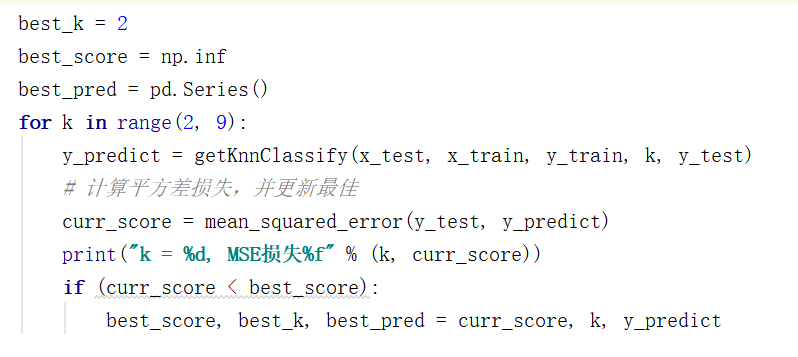


其中，x是包含了LRMC这4个指标的4列，y是F指标的0-1值（即上述是否为“常客”的两列）

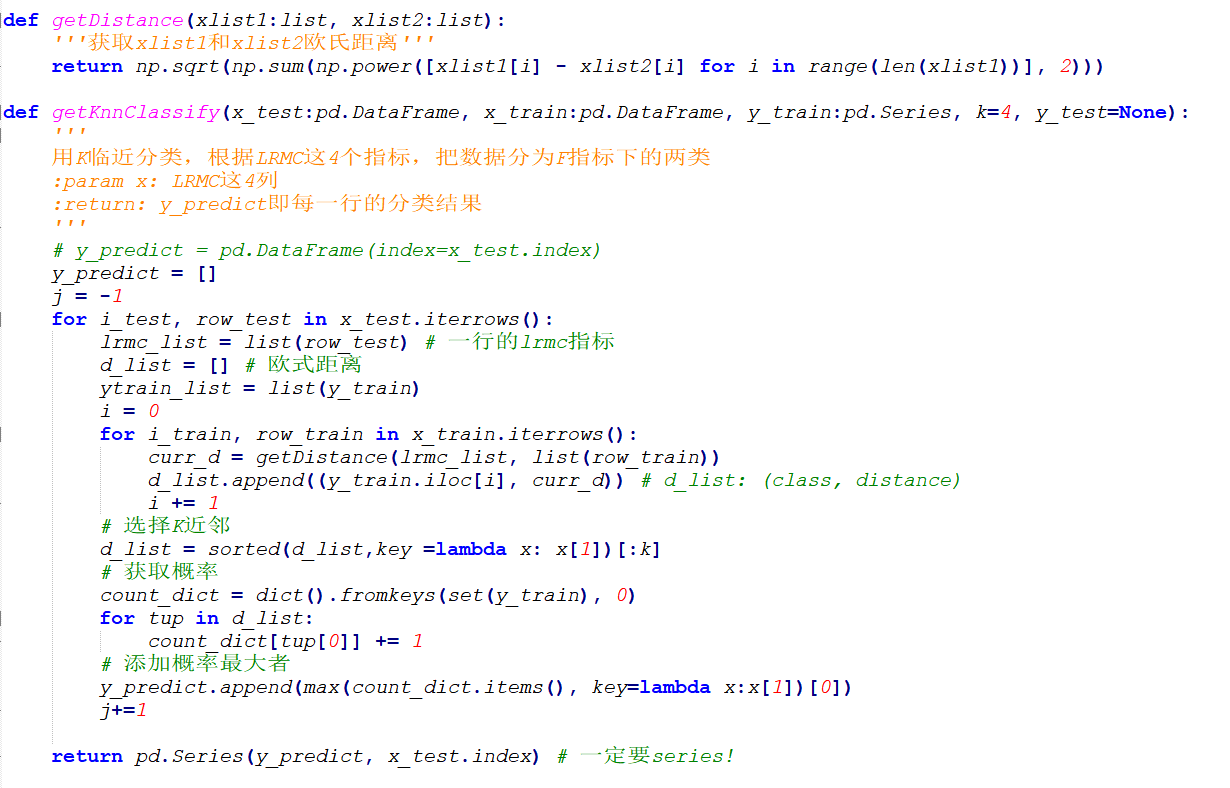
（2）把数据随机划分为训练集和测试集：



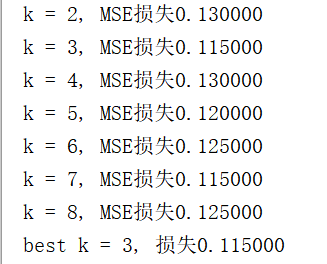
（3）尝试取不同的K值，用K近邻法，对测试集进行分类



getKnnClassify函数见下页

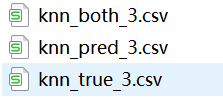


**5、实验结果**

不同K值下的损失：

可见最佳的K是3，即挑选临近的3个训练集的点。

预测结果保存在下面3个文件夹，knn\_both第一列测试集预测结果，第二列测试集真实；knn\_pred前几列分别为除了F外的4个指标，最后一列是测试集预测结果



**6、实验总结**

有了上个实验的积累，分类的实现就好做多了，而且KNN比apriori算法还简单一点。虽然实现上面仍然踩了一些坑，毕竟之前的确很久没用pandas这些了，但后来问题也解决了。

**暨南大学本科实验报告专用纸**

课程名称 数据挖掘 成绩评定

实验名称 实验三 聚类 指导教师 刘波

试验编号 0806011603 类型 综合性

学生姓名 邝庆璇 学号 2016051598

学院 信息科学技术学院 系 计算机 专业 计算机科学与技术

**1、实验目的：**

（1）理解聚类算法原理

（2）掌握并熟悉聚类算法的使用

**2、实验内容**

利用K近邻算法，根据实验一中预处理得到的LRMC这5个指标，把乘客进行聚类。

**3、实验原理**

原理如书上所言。

1.创建k个点作为起始质心，可以随机选择(位于数据边界内)；

2.while 任意一个点的簇分配结果发生改变：

　　3.For 数据集中每一个点：

　　　　①对每个质心计算质心与数据点之间的距离

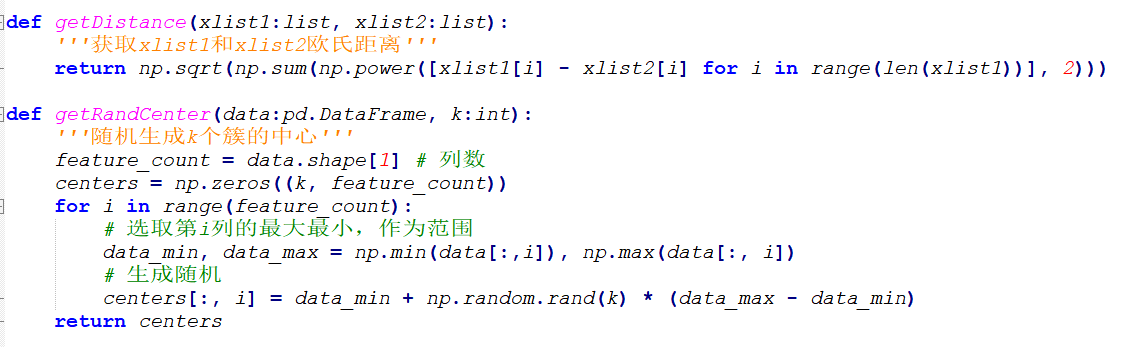
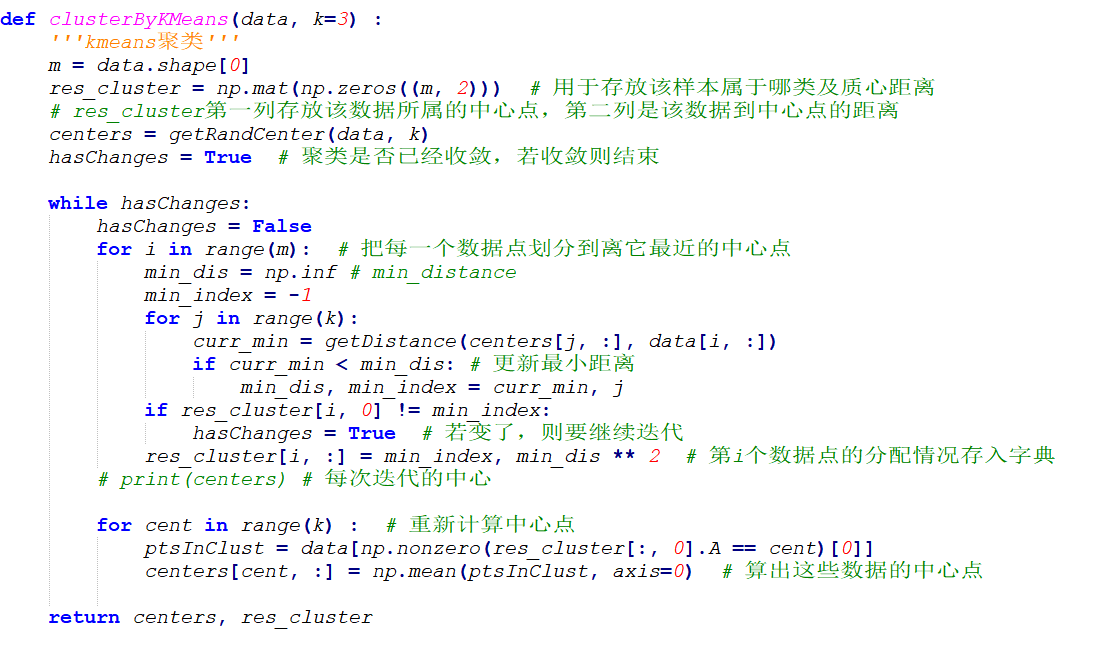
　　　　②将数据点分配到距其最近的簇

　　4.对每一个簇，计算簇中所有点的均值并将均值作为质心；回到2；

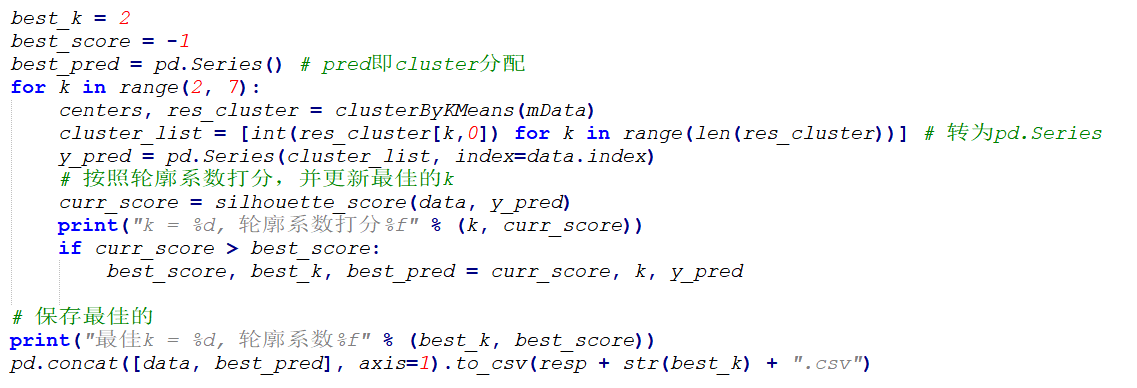
**4、实验步骤**

尝试K从2到6，对KRMFC这5个指标进行聚类

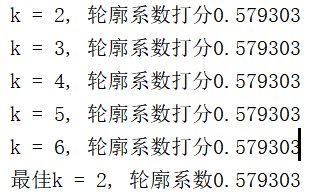
核心代码如下：



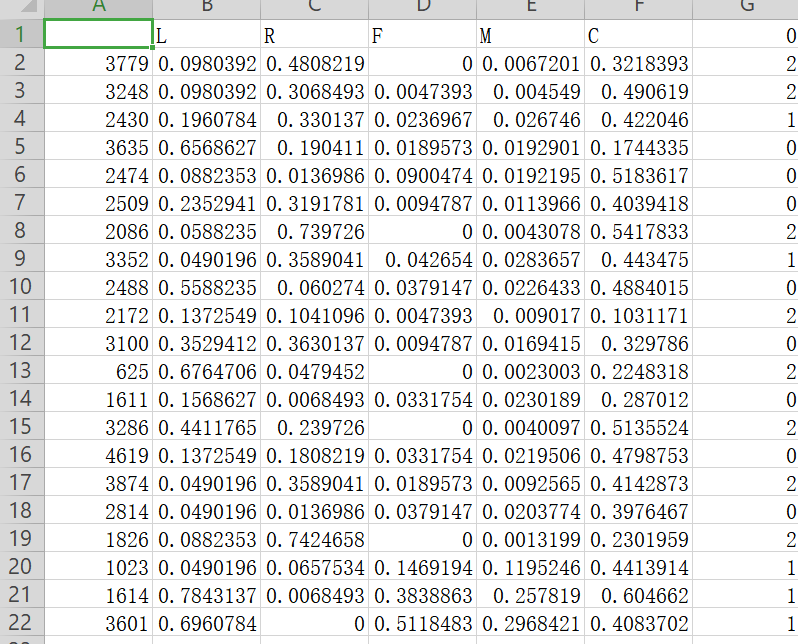
利用轮廓系数来对聚类效果打分，选取最佳效果的K



**5、实验结果**



得分比较接近，可能是这5个特征的数据并不合适聚类，也许需要进行降维或者其他处理。结果保存在kmeans\_result\_3.csv这个文件中，截图如下：



（最后一列为类别）

**6、实验总结**

KMeans聚类的思想不难，但这次想试试不用pandas的DataFrame，改用了numpy的np.matrix，有不少不熟悉的地方，但也学到了不少。