《大数据分析方法》课程实验报告

学号： 2020204246

姓名： 王浩

专业： 计算机科学与技术图灵班

班级： 2020级

# 实验一：关联规则算法实验

**实验目的：**在掌握关联规则算法的原理的基础上，能够应用关联规则算法解决实际问题。

**实验内容：**根据实验数据，采用Apriori等关联规则发现算法，给出相关关联规则。

**实验要求：**给出数据预处理过程、关联规则发现算法及发现关联规则，并对关联规则结果进行分析说明。

**实验题目：**蔬菜价格相关性分析

蔬菜的价格会受季节、天气等多方面因素的影响，但许多会出现同涨或者同跌等现象，请根据给出的蔬菜价格数据，采用关联规则发现算法，发现哪些蔬菜之间具有同涨、同跌或者涨跌不同步的现象。

数据格式如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 日期 | 蔬菜名 | 价格 | 肉食禽蛋 | 批发价格 |
| 2009/6/11 | 大白菜 | 0.7 | 精猪肚 | 28 |
| 2009/6/11 | 青萝卜 | 1.2 | 猪口条 | 28 |
| 2009/6/11 | 胡萝卜 | 2 | 猪心 | 24 |
| 2009/6/11 | 水萝卜 |  | 猪肝 | 7 |
| 2009/6/11 | 葱 | 2.5 | 猪大肠 | 22 |
| 2009/6/11 | 芽葱 |  | 猪蹄 | 15 |
| 2009/6/11 | 元葱 | 0.8 | 猪大排 | 20 |
| 2009/6/11 | 生姜 | 3 | 猪肋排 | 25 |
| 2009/6/11 | 大蒜 | 3 | 精猪肉 | 17 |
| 2009/6/11 | 土豆 | 1.3 | 精牛肉 | 30 |
| 2009/6/11 | 菠菜 | 1.4 | 白条肉 | 16 |
| 2009/6/11 | 韭菜 | 1 | 羊肉片 | 32 |
| 2009/6/11 | 芹菜 | 1.5 | 西装鸡 | 12 |
| 2009/6/11 | 油菜 | 1 | 鸡大腿 | 12 |
| 2009/6/11 | 苔菜 |  | 鸡翅 | 18 |
| 2009/6/11 | 香菜 | 2 | 鸡爪 | 16 |
| 2009/6/11 | 大头菜 | 0.6 | 鸡胸肉 | 14 |
| 2009/6/11 | 芸豆 | 1.2 | 翅根 | 14 |
| 2009/6/11 | 豆角 | 2 | 翅中 | 26.2 |
| 2009/6/11 | 茄子 | 1.2 |  |  |
| 2009/6/11 | 青椒 | 1.7 | 鸡心 | 11 |
| 2009/6/11 | 西红柿 | 1.4 | 松花蛋 |  |
| 2009/6/11 | 黄瓜 | 1.1 | 咸鸭蛋 |  |
| 2009/6/11 | 茭瓜 | 0.7 | 鸡蛋 | 6.6 |
| 2009/6/11 | 冬瓜 | 1.6 |  |  |

**实验过程：**

1. **数据预处理**

**因为excel表格中的数据并未按照日期顺序排序，索引先通过excel中的排序功能对日期进行排序，然后遍历原始表格，创建新表。**

**发现规律，每个日期占有的行数相同，且相邻日期之间有6个空行，**

**相同日期间各个物品的相对位置固定。**

**对数据进行处理以日期为行，每个物品为列，遍历填值。**

**原始数据格式： 经过过渡处理后的数据：**

一些文字和图片的手机截图

描述已自动生成 

**最终的数据形式：**

图形用户界面, 表格

描述已自动生成

1. **Apriori算法发现关联规则**

**首先对所有数据，进行apriori算法，求出频繁项集**

**# 设置数据集**

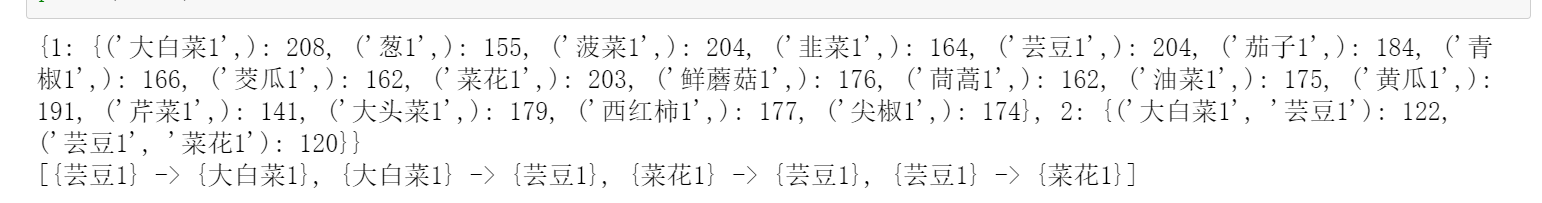
**# 挖掘频繁项集和频繁规则**

**itemsets, rules = apriori(data\_list, min\_support=0.25, min\_confidence=0.5)**

**print(itemsets)**

**print(rules)**

**实验结果：**



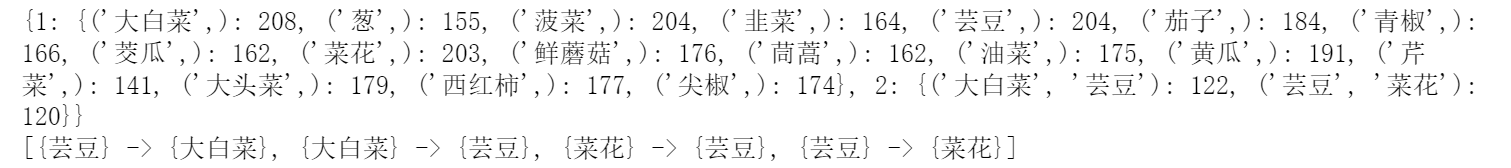
**因为肉禽蛋奶中有很多缺失值大，而蔬菜中缺失值较少，两者加在一起导致总事务过多，结果最小支持度过低，找出的频繁项集数量少。**

**所以我们又采取对蔬菜和肉类分开进行关联性分析的方案，分别构建蔬菜价格信息表和肉类价格信息表，蔬菜间建立蔬菜关联规则，肉类间建立肉类关联规则，与一开始总类之间进行对比。**

**蔬菜类：**

**处理代码：**

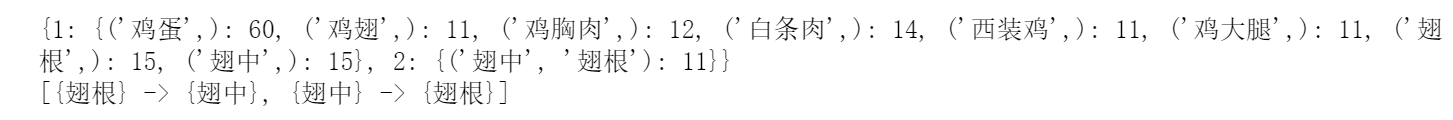
**实验结果：**



**肉禽蛋奶类：**

**处理代码：**

**实验结果：**



通过实验结果我们可以发现：

通过多次调整最小支持度和最小置信度的参数

芸豆和大白菜，菜花和芸豆，翅根和翅中，精猪肉和白条肉，芸豆和油菜，尖椒和大白菜之间具有较大相关性。

**实验思考：**

通过蔬菜价格关联性分析的实验加深了对于Apriori算法的理解，在此之前能够通过题目理解Apriori算法并解答；实验过程是用代码实现这一解题过程，能够对解题过程理解更加透彻，并且锻炼了编程能力；

Apriori算法的相关属性定理：

如果项集X 是频繁项集，那么它的所有非空子集都是频繁项集。

如果项集X 是非频繁项集，那么它的所有超集都是非频繁项集。

Apriori算法的性能瓶颈：

多次扫描事务数据库，需要很大的I/O负载；

可能产生庞大的候选集；

Apriori算法的改进方法：

事务压缩、划分、选样、动态项集计

实验设计的内容主要有数据预处理、关联规则发现Apriori算法，关联规则结果的分析说明等；通过查阅资料理解程序的结构，并且学会了一些函数的使用方法，丰富了自己的知识。

# 实验二：回归算法的实验

**实验目的：**在学习线性回归、决策树、贝叶斯分类、神经网络等算法的原理的基础上，掌握已有的数据分类算法软件包中数据分类回归算法的使用。

**实验内容：**学习scikit-learn包中的[Classification](http://scikit-learn.org/stable/supervised_learning.html" \l "supervised-learning)，网络地址：http://scikit-learn.org/

**实验要求：**给出数据处理过程说明，算法预测过程，结果说明

**实验题目：美国新冠肺炎总确诊量预测**

今年，新冠肺炎成为全球重大公共卫生事件，美国是当前疫情最严重的国家，其感染人数与死亡人数居世界首位。目前以获取从2020.1.28-2020.9.9之间的数据，请利用所学知识，采用2020.1.28-2020.8.31的数据建立模型，并用9.1-9.9数据进行模型测试，说明模型的有效性。

数据格式如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **时间** | **新增** | **总确诊** | **治愈** | **死亡** |
| 2020.1.28 | 0 | 5 | 0 | 0 |
| 2020.1.29 | 0 | 5 | 0 | 0 |
| 2020.1.30 | 1 | 6 | 0 | 0 |
| 2020.1.31 | 0 | 6 | 0 | 0 |
| 2020.2.01 | 1 | 7 | 0 | 0 |
| 2020.2.02 | 1 | 8 | 0 | 0 |
| 2020.2.03 | 3 | 11 | 1 | 0 |
| 2020.2.04 | 0 | 11 | 1 | 0 |
| 2020.2.05 | 0 | 11 | 1 | 0 |
| 2020.2.06 | 1 | 12 | 1 | 0 |
| 2020.2.07 | 0 | 12 | 1 | 0 |
| 2020.2.08 | 0 | 12 | 1 | 0 |
| 2020.2.09 | 0 | 12 | 1 | 0 |
| 2020.2.10 | 0 | 12 | 1 | 0 |
| 2020.2.11 | 1 | 13 | 1 | 0 |
| 2020.2.12 | 0 | 13 | 3 | 0 |
| 2020.2.13 | 1 | 14 | 3 | 0 |
| 2020.2.14 | 1 | 15 | 3 | 0 |
| 2020.2.15 | 0 | 15 | 3 | 0 |
| 2020.2.16 | 0 | 15 | 3 | 0 |
| 2020.2.17 | 0 | 15 | 3 | 0 |

## 实验目标:

根据过去共226条数据信息，预测未来新冠肺炎总确诊人数，

假设每天的总确诊人数为，模型参数为，预测总确诊人数。

## 实验分析

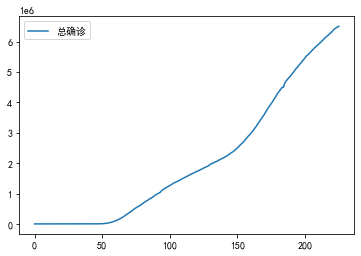
因为通过当前时间的数据来拟合没有实际意义，需要采用之前的数据来对之后的数据进行预测，而时间序列模型能够根据时间顺序，通过前几天的数据预测后一天的数据，因此我采用时间序列预测模型ARMA模型对总确诊人数进行预测。

## 实验过程及部分结果：

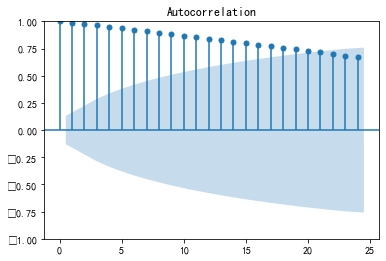
1. 首先对数据进行分析，
2. 然后通过分析数据特征选择ARMA模型，
3. 其次确定模型参数，
4. 最后训练模型并进行检验

### 数据分析

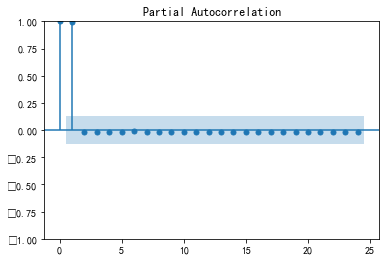
分别绘制时序图，自相关系数，偏自相关系数，观察数据的特征和平稳性，



时序图

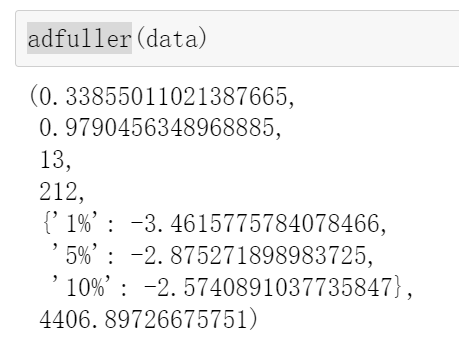


自相关系数



偏自相关系数

进行单位根检验平稳性



经过分析以上检验结果，我发现因为数据本身递增，不具备平稳性的特征。

因此，即便把数据进行差分处理平稳化，结果得到的数据，仍不具有平稳性。

所以选择ARIMA模型。

### 确定参数

之后首先通过'aic','bic','hqic暴力找到ARMA模型的参数。

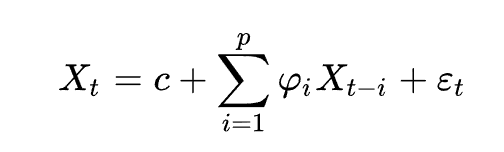


### 训练ARMA模型

调用库函数进行训练

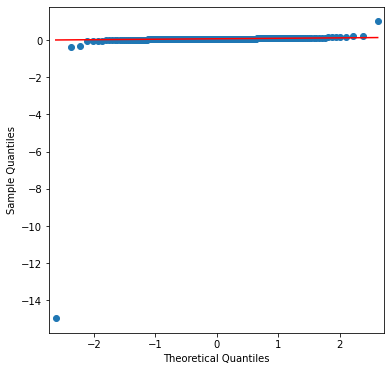
采用ARMA中的AR模型，因为ARMA是AR模型和MA模型的合成，

在本实验中主要采用AR模型，下图为AR模型数学原理：



### 模型检测

绘制qq图



根据上图，说明数据不具有正太分布特征。

DW检验

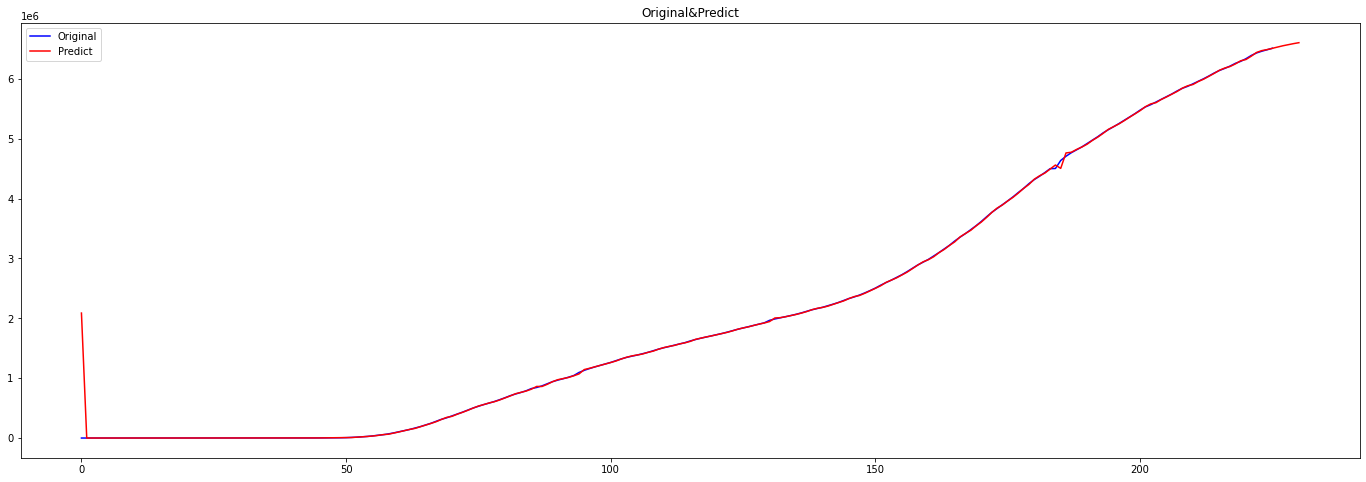


说明数据不具有独立性。

预测未来一天的总确诊人数



将时间序列结果与实际结果进行拟合



预测值与真实值的图像

数据评估：



## 心得体会

一开始选择审题出现了问题，设想得是通过另外三列来拟合预测，总确诊人数这一列，

尝试进行线性回归，多项式回归以及xgboost机器学习程序进行拟合，最后发现xgboost能够很好地拟合数据。

最后发现这样做不具有实际意义，预测肯定不能用到当天的数据。有点可惜试了很多方法，以为做完实验了，结果发现一开始的方向就错了，不过在过程中学到了很多回归以及机器学习的相关知识。

随后采用时间序列模型。

时间序列模型不一定非得要求数据平稳，需要看模型的原理，比如AR模型，就是根据前几天的数据来预测之后的数据并不要求数据平稳性，而MA模型要处理的数据是震荡的，有相同的特征，需要数据平稳化。

而原数据就是不平稳的，经过差分后可能还是不平稳，所以模型选择上应该采用AR模型。最后拟合效果很不错。

时间序列模型可以用在数据预处理上，时间序列是只能在一列上进行操作。

# 实验三：乳腺癌决策树算法实验

**实验目的：**在掌握决策树相关算法的原理的基础上，能够应用决策树算法解决实际问题。

**实验内容：**根据实验数据，采用决策树算法，给出分类结果，并评估模型准确性。

**实验要求：**给出数据预处理过程、决策树算法，并对分类结果进行分析说明。

**实验题目：**乳腺癌决策树算法

## 实验目标:

根据过去共569条数据信息，预测肿瘤是良性还是恶性。

## 实验分析

这是一个有监督的分类问题

采用决策树算法进行，有监督学习，实现分类，并预测。

首先直接使用决策树算法，进行预测，并进行模型评估，

然后通过模型，计算出决策树算法的最优参数，比较模型评分之间的差异

最后画出决策树算法的分类结果图，进行直观了解决策树算法的分类结果。

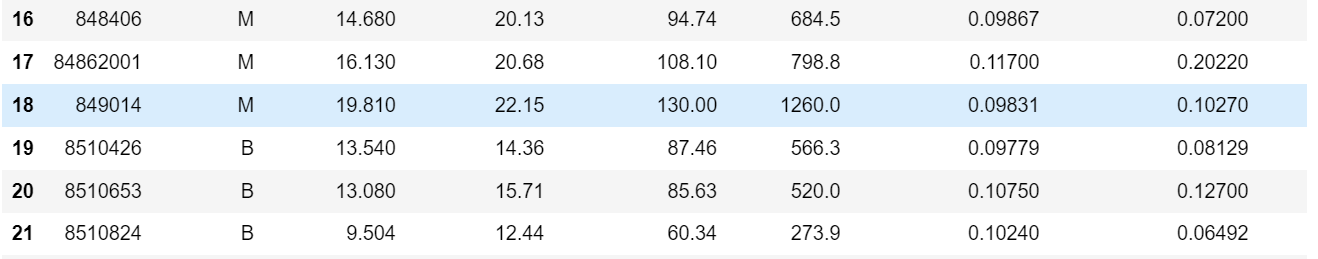
## 实验过程及结果：

1. 首先对数据进行分析，发现数据没有缺失值，并且格式比较规整，因此预处理

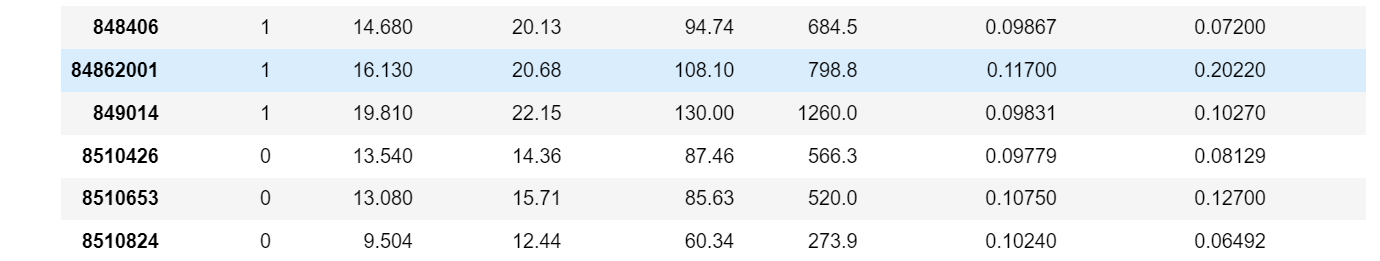
只提取了标签和特征。

1. 然后通过分析数据特征结果选择决策树模型，训练并验证。
2. 通过GridSearchCV库确定模型最优参数，
3. 最后训练模型并进行检验
4. 画出决策树分类效果图

### 数据分析



原始数据



预处理后数据

### 模型训练并评估

通过调用决策树进行模型训练，并评估



### 通过模型选择最优参数并评估检验

训练结果如下

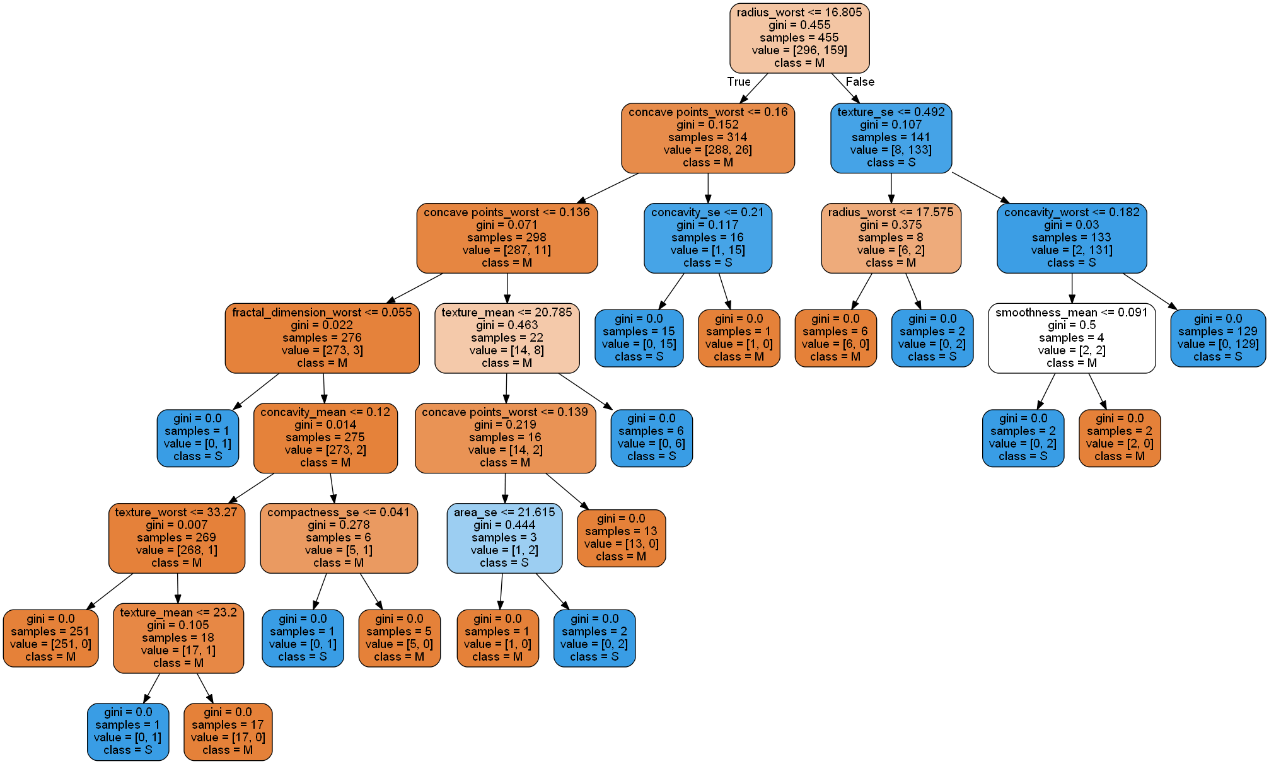




预测结果随模型参数调整变化图

### 画出决策树分类图

导入Graphviz库-画决策树图



决策树分类图

## 心得体会

采用决策树模型，训练并验证，通过GridSearchCV库确定模型最优参数，训练模型并进行检验。在这一过程中感受到了决策树算法的原理，分类依据，同时熟悉了基本的数据分析算法流程。

学会了通过导入Graphviz库-画决策树图等一系列效果图，通过观察感受决策树的分类依据和原理。

# 实验四 分类算法应用实验

**实验目的：**在掌握分类算法基础实验应用的基础上，能够应用分类算法解决实际问题。

**实验内容：**根据实验数据，采用深度学习算法，解决实际应用问题。

**实验要求：**给出实际问题的解决方案，数据预处理过程、分类模型的建立过程、参数调优、结果分析实验题目：乳腺癌预测

**实验题目：**猫狗大战

问题源于Kaggle(www.kaggle.com)猫狗图片识别比赛，数据集中共包含猫、狗的图片各12500张，要求设计一个分类算法，实现对图片为“猫”或者是“狗”的图片的分类。

数据集下载地址：https://www.kaggle.com/c/dogs-vs-cats/data

## 一.问题描述

通过Alex卷积神经网络对Cat-Dog图片数据集进行学习，建立分类模型，通过模型预测图片类别。

## 二.解决思路

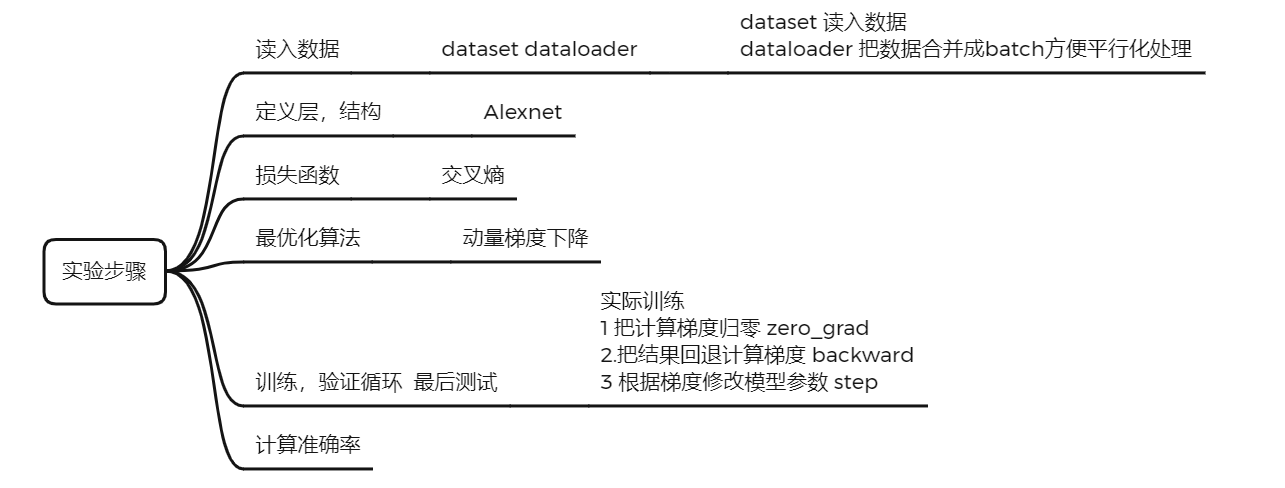


图1 问题解决流程图

如上述流程图所示，首先读入数据，将图片转化为tensor，根据Alexnet卷积神经网络示意图，定义网络结构，选择合适的损失函数和最优化算法。之后对数据进行训练，本模型训练了10epoch，最后进行准确率评判。

## 三.解决步骤

### 3.1 数据读入

#### 3.1.1 数据预处理

将图片大小转化为227\*227，转化类型为tensor，并进行归一化。



图2 数据预处理函数

#### 3.1.2 数据集划分

将数据集划分为训练集和测试集，取猫狗各500张图片作为测试集，其余图片为训练集。

### 3.2 定义神经网络结构

#### 3.2.1 LRN

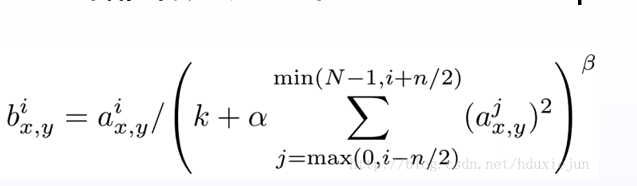


图3 LRN层神经网络

根据上图所示为LRN层计算公式，写出如下LRN层代码

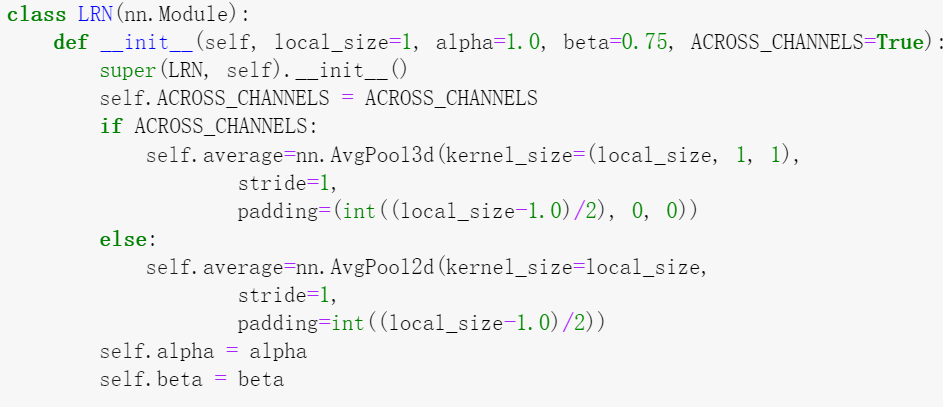


图4 LRN层神经网络代码

#### 3.2.2 ALEX

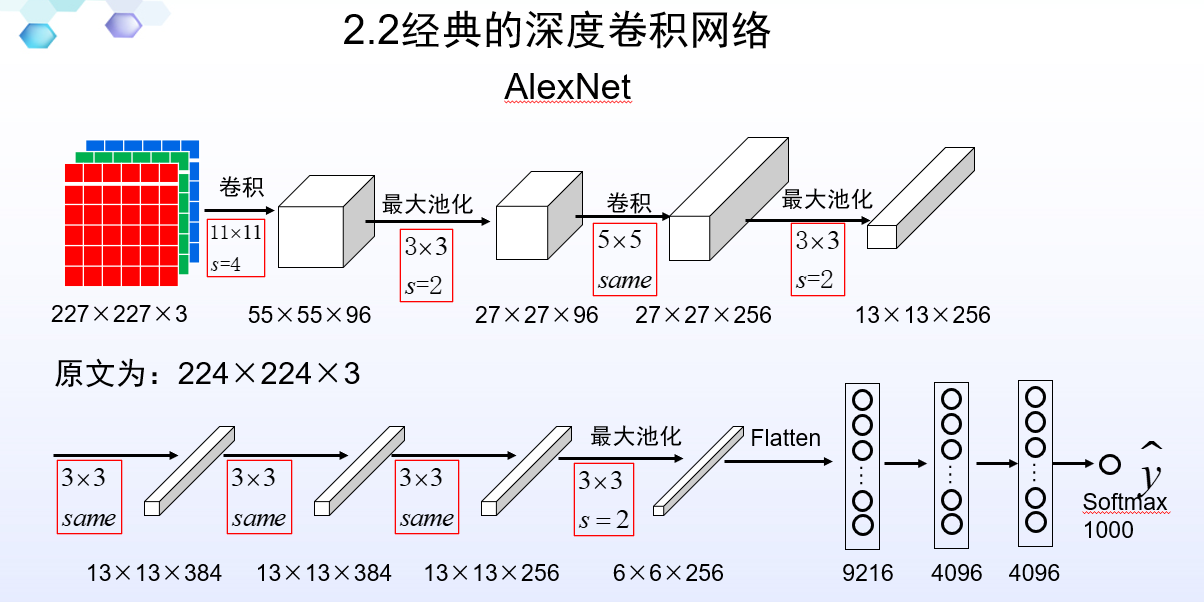


图5 ALEX神经网络结构

根据如上图所示的ALEX神经网络结构，定义如下ALEX神经网络结构，核心代码如下：

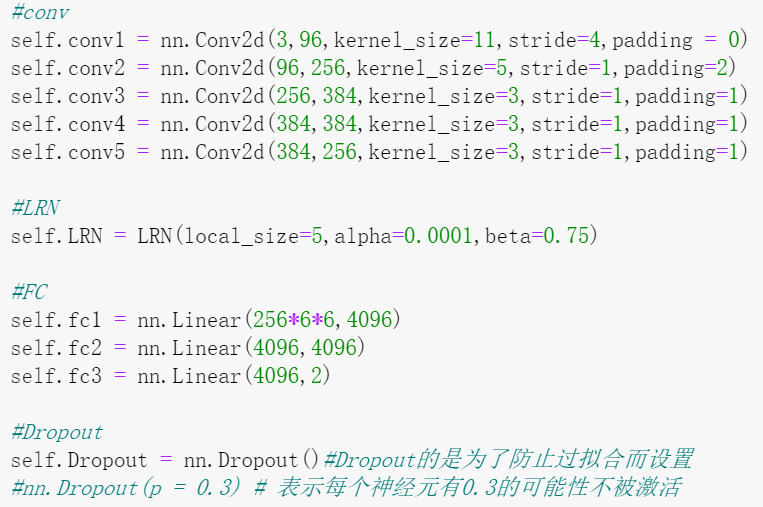


图6 ALEX神经网络代码

为防止过拟合，采用Dropout函数，随机30%的神经元不被激活，提高模型的健壮性和准确性。

#### 3.2.3 定义损失函数



图7 定义损失函数代码

本实验采用交叉熵作为损失函数，因为本问题是分类问题，采用交叉熵比较合适。

交叉熵计算公式如下所示

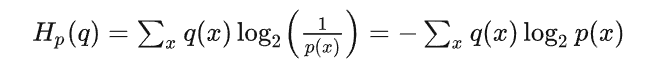


图8 交叉熵计算公式

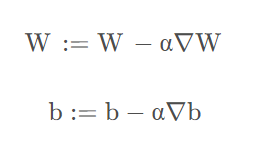
#### 3.2.4 选择优化器



图9 定义优化器代码

采用动量梯度下降方法，尽可能防止模型陷入局部最小值而不能到达全局最小值点。

动量梯度下降法的公式如下：



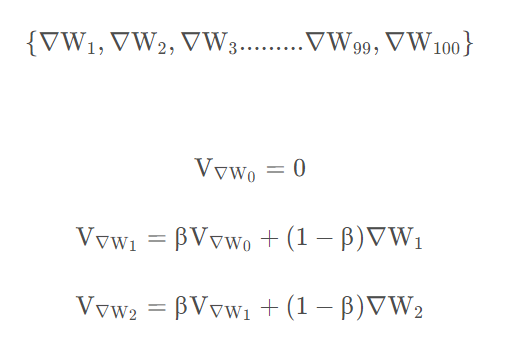


图10 动量梯度公式

动量梯度下降法则对各个mini-batch求得的梯度,利用

使用指数加权平均得到 ，同时采用学习率a以及参数β，来更新之前的参数。

### 3.3 模型训练

#### 3.3.1 train\_loop

步骤：

1 把计算梯度归零 zero\_grad

2.把结果回退计算梯度 backward

3 根据梯度修改模型参数 step



图11 训练模型代码

训练过程代码输出



图12 训练过程中间结果

#### 3.3.2 计算准确率

准确率计算方法设为预测值与实际值相等的数量，为总测试集的数量，acc为准确率，计算公式如下



图13 准确率计算公式



图14 计算准确率代码

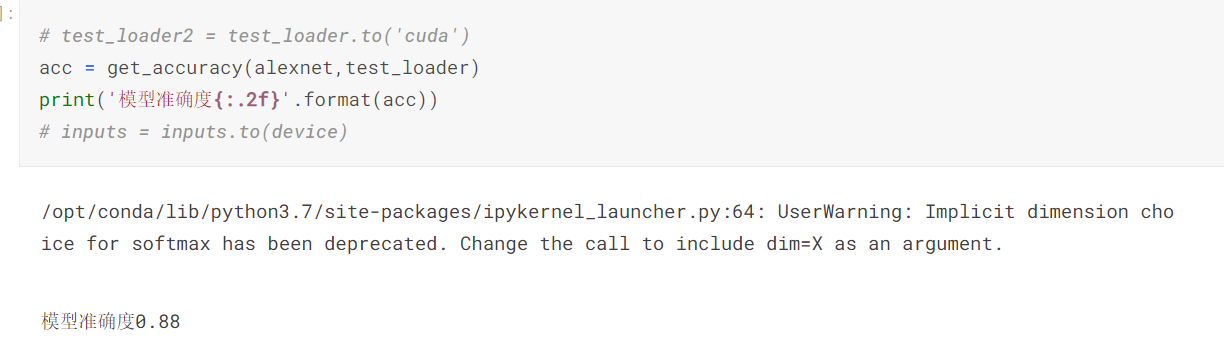


图15 准确率结果



图16 损失率随epoch变化图

# 实验五 聚类算法实验

**实验目的：**在掌握聚类算法基础原理基础上，掌握应用聚类算法解决实际问题。

**实验内容：**根据给定的实验数据，采用层次聚类、基于划分的聚类、密度聚类等算法对数据聚类。

**实验要求：**给出实际问题的解决方案，数据预处理过程、聚类算法的建立过程、结果分析。

**实验题目：**《红楼梦》作者分析

《红楼梦》是我国著名的四大名著之一，一般的认为《红楼梦》的前八十回为曹雪芹撰写，后四十回为高鹗续写，但也有学者对此并不认可。能否利用聚类分析的方法来对《红楼梦》的作者进行分析。

提示思路：一般来说，不同的作者往往会具有不同的写作风格，这些风格可以通过在文中的虚词的频率进行衡量，因此，可以考虑统计各章中虚词出现频率，并以此作为基础数据来聚类分析，对《红楼梦》章节进行划分，从而分析章节与作者之间的关系。

数据：《红楼梦》完整版.txt

参考文献：李贤平.《红楼梦》成书新说.复旦学报(社会科学版).1987年第5期.

## 问题描述

通过K-means聚类模型以及层次聚类模型对红楼梦中每一回各个虚词出现的频率进行聚类分析，预测作者数量。

## 解决思路

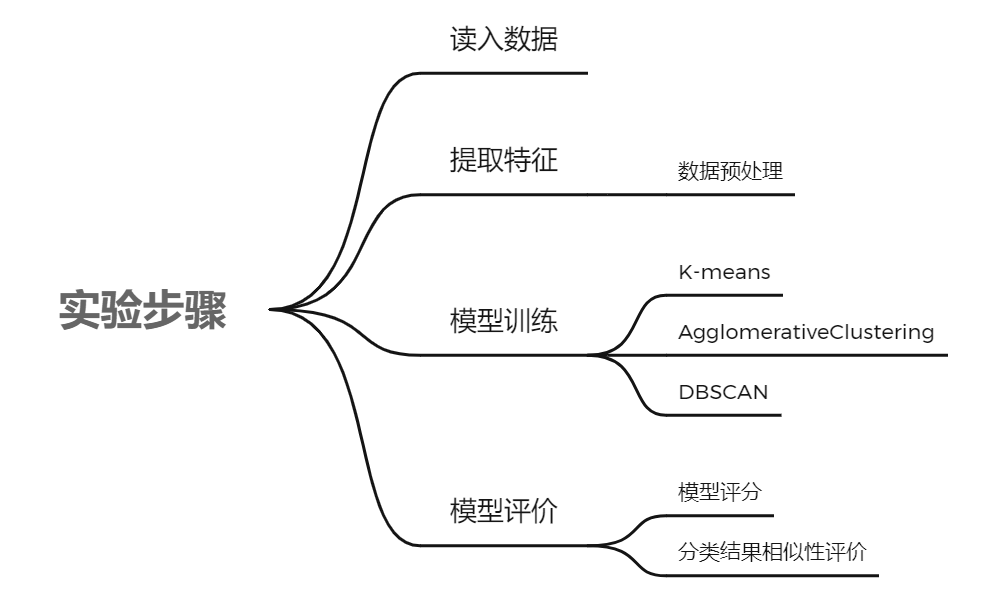


图1 问题解决流程图

通过阅读论文以及对虚词的分析，找到需要研究的对象46个虚词，然后如上述流程图所示，首先按行读入数据，通过正则表达式，找到每一回开始的行数，然后计算每一回中各个虚词的出现频率，作为分类特征，带入K-means，AgglomerativeClustering，DBSCAN模型中进行模型训练，评分，最后对分类结果的相似性进行评价。

## 解决步骤

### 3.1 数据读入

#### 3.1.1 数据预处理

先确定每一回的区间，

通过正则表达式找到所有的回开始的行号

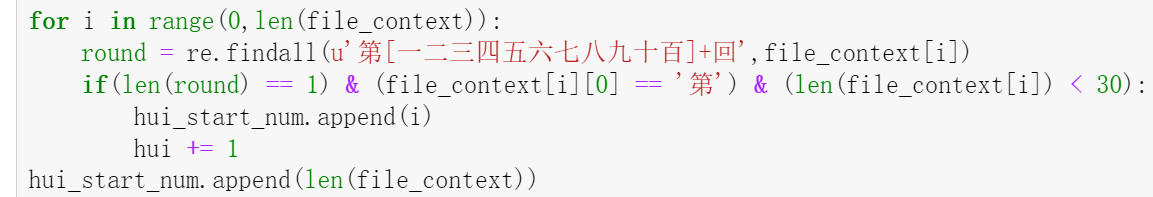


图1 确定回开始行号代码

统计每个虚词在每一回出现的频率，首先利用jieba库进行分词，统计出总词数，然后通过正则表达式统计出虚词出现的次数，因为直接相除频率过小，就将每一个频率扩大100倍，方便后续处理。

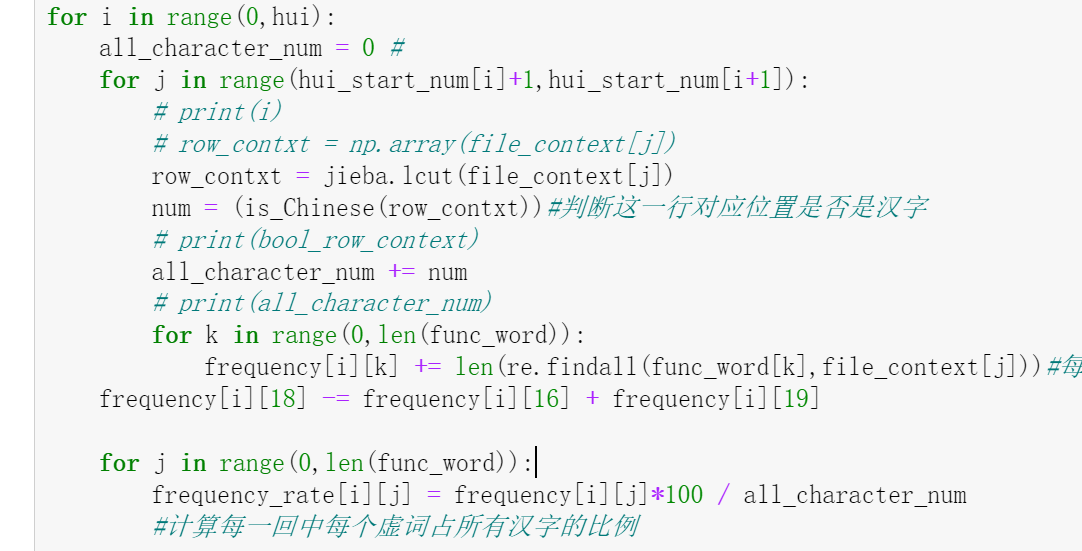


图2 提取特征核心代码

### 3.2 模型的构建与预测

#### 3.2.1 基于划分聚类 K-means

K-Means算法主要是根据两簇之间的距离，将距离近的两簇合并，使簇内距离尽可能小，簇间距离尽可能大。

**通过绘制DBI得分曲线**，根据曲线变化确定最佳分类数

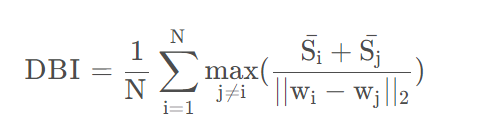


图3 DBI计算公式

DBI的值最小是0，值越小，代表聚类效果越好

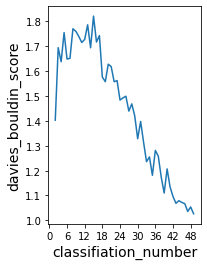


图4 DBI评分随分类数量变化图像

**绘制CH得分曲线**，根据曲线变化确定最佳分类数

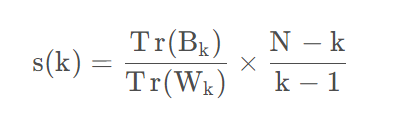


图5 CH计算公式

得到的Calinski-Harabasz分数值越大则聚类效果越好。

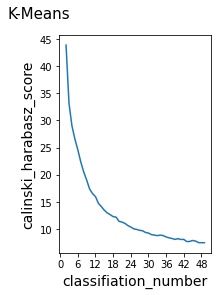


图6 CH随分类数量变化图像

核心代码：

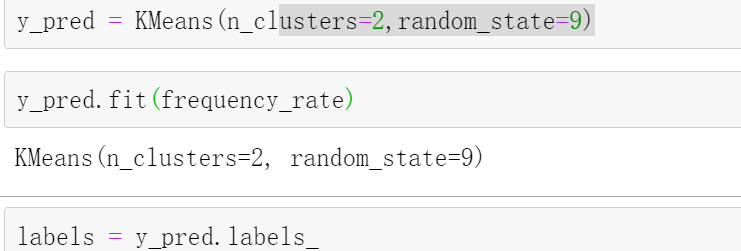


图7 K-means核心代码

#### 3.2.2 基于层次聚类 AgglomerativeClustering

层次聚类的算法主要是将样本自下而上，逐步合并或是自顶向下，将样本分解。计算聚类簇间距离的方法有三种，分别为Single Linkage，Complete Linkage和Average Linkage。

对比采用不同的linkage和不同的分类数对分类效果的影响，选取分类得分最高的作为分类算法

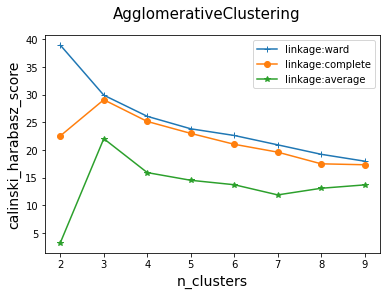


图8 **不同linkage下模型分类评分随分类数量的变化图**

根据如上图所示ward linkage方式聚类效果另外两种分类效果好，因此选取ward作为linkage

#### 3.2.3 基于密度聚类 DBSCAN

DBSCAN是一种基于密度的聚类算法，通过设置eps，min\_samples等参数进行聚类，因此在使用DBSCAN算法进行聚类时主要问题就是确定这两个参数的值。

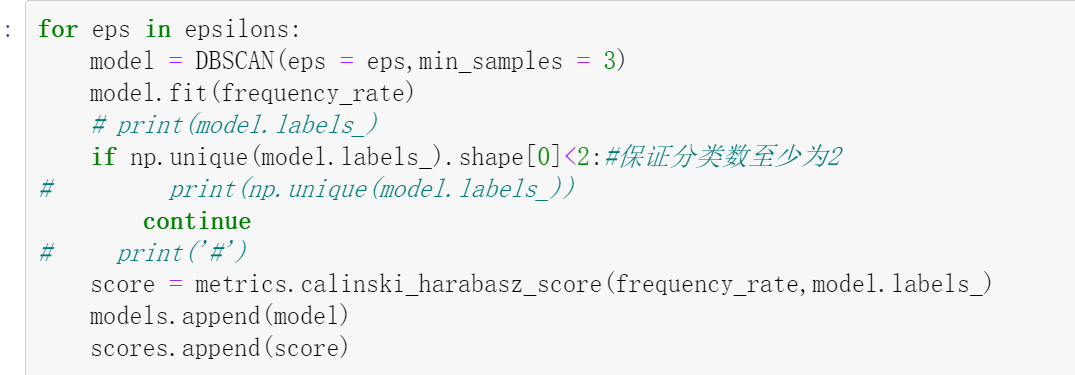


图9 枚举参数代码

## 四. 结果分析

#### 3.3.1 模型分类结果

K-means分类结果

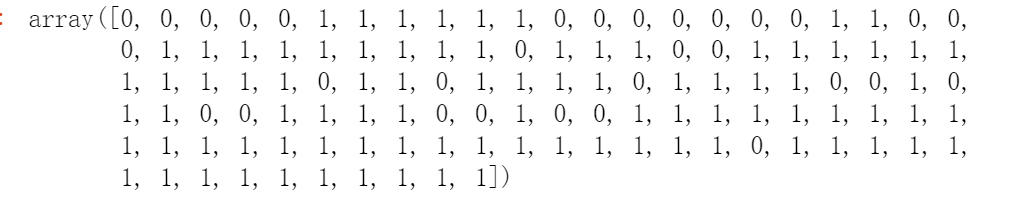


图10 K-means聚类标签结果

AgglomerativeClustering**模型分类结果**

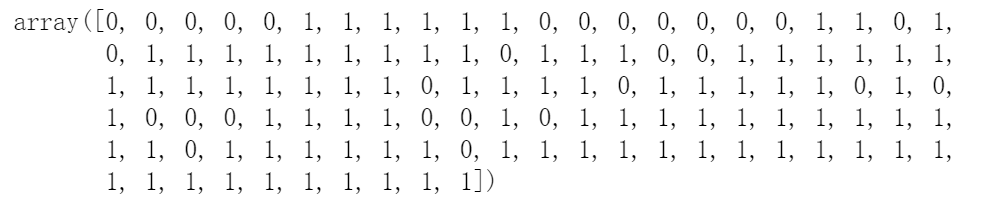


图11 AgglomerativeClustering**模型分类结果图**

**DBSCAN模型分类结果**



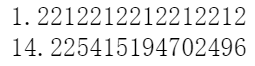


图12 DBSCAN模型评分图

#### 3.3.2 模型分类结果相似度

将K-means 分类结果和层次聚类结果进行相似度比较

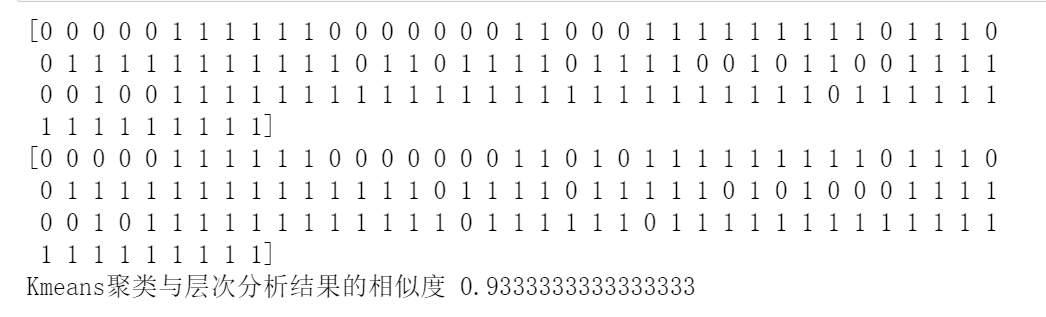


图14 模型分类结果相似度图