**北京邮电大学软件学院**

**2019-2020 学年第 1学期实验报告**

**课程名称： 数据挖掘**

**实验名称： 实验一：weka使用与聚类分析**

**实验完成人：**

**姓名： 平雅霓 学号： 2017211949**

**指导教师： 牛琨**

**日 期： 2019 年 10 月 26 日**

1. **实验目的**

熟悉WEKA软件的使用，加深对数据处理和聚类分析的理解。

1. **实验内容**
2. 熟悉WEKA软件的使用；
3. 给出本数据的数据描述报告；
4. 采用K均值算法对给定数据集进行聚类，给出聚类结果。改变K的不同取值，研究K值改变给聚类结果所带来的变化。改变初始簇心，研究簇心变化给聚类结果所带来的变化；
5. 给出你认为最优的模型并加以解释。
6. **实验环境**

Windows环境、weka-3-8

1. **实验过程及结果**
2. **数据描述**

该数据集收集了2003年-2012年之间上市的1411款手机的相关信息，其中的每条数据项中属性包括产品型号、品牌、颜色、价格、市场定位、硬件等。

● 产品ID：自增长，从1开始，到1411.

● 产品颜色: 共有7个颜色，众数为1。

● 产品上市时间：共有9个年限，众数为2010年。

● 产品市场定位：共有7个定位，众数为0（即经济实用型）。

● 芯片平台：共有高通和威盛两个值，众数为高通。

● G网：众数为0，即大多数为G网。

● 芯片主频：取值在40~2400之间，平均数为196.149，众数为96。

● AP：取值为0有830个，1有581个，即众数为0。

● 频段数量：有5个取值，众数为1。

● 零售价格：取值为184~9380，平均值为1117.0198，众数为298。

● 外观类型：有5个取值，其中众数为0，占总数的82.1%。

● 厚度：取值为9~85，平均厚度为14.9489，众数为15。

● 产品重量：范围为48.4~790.2，平均重量为107.6030，众数为110。

● 屏幕数量：有单双之分，众数为1，即大多数为95.2%为单屏幕。

● 主屏幕尺寸：取值为0~7，平均值为2.41，众数为2.4。

● 显示分辩率：平均值为82613.703，众数为76800。

● 触摸屏：有三种触摸屏，众数为0，即大多数为为触摸屏。

● 键盘类型：有三种键盘，众数为1，即大多数为数字键盘。

● RAM：取值范围为1~4096，平均值为139.4232，众数为64。

● ROM：取值范围为0~16384，平均值为306.5258，众数为128。

● Flash内存：取值范围为0~16384，平均值为239.5981，众数为0。

● 摄像头：取值范围为0~1300，平均值为125，众数为30。

● 定位：众数为0，即71.2%无定位功能。

● FM广播：众数为0，即53.8%无FM广播。

● 电视：众数为0，即98.6%无点视功能。

● Modem：众数为1。即60%有调制解调器。

● 红外：众数为0，即98.6%无红外。

● 蓝牙：众数为0，55%无蓝牙功能，45%有蓝牙功能。

● WLAN：众数为0，86.9%手机不能使用无线网。

● 电池容量：平均数为1134.8936，众数为1000。

● 重力感应器：众数为0，82.4%的手机无重力感应器。

● 方向感应器：众数为0，90.7%的手机无方向感应器。

● 文字输入方法数：有3种输入法，众数为1。

● 智能系统：众数为0，81.7%的手机无智能系统。

详细的属性描述如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 产品ID | 产品编号，每个产品的唯一标识，1~1411 | 触摸屏 | 无0,电阻1,电容2 |
| 产品型号 | 每个产品所属型号 | 键盘类型 | 无0,数字1,全2 |
| 产品品牌 | 每个产品所属品牌 | RAM | 1~4096 |
| 产品颜色 | 1~7 | ROM | 0~16384 |
| 产品上市时间 | 2003~2012 | Flash内存 | 0~16384 |
| 产品市场定位 | 经济实用0,新潮炫酷1,娱乐小资2,商务功能3  品味尊贵4,老年机5,儿童机6 | 摄像头 | 0~1300 |
| 芯片平台 | 高通0,威盛1 | 定位 | 是1,否0 |
| G网 | 是1,否0 | FM广播 | 是1,否0 |
| 芯片主频 | 40~2400 | 电视 | 是1,否0 |
| AP | 是1,否0 | Modem | 是1,否0 |
| 频段数量 | 1~5 | 红外 | 是1,否0 |
| 零售价格 | 184~9380 | 蓝牙 | 是1,否0 |
| 外观类型 | 直板0,翻盖1,滑盖2,旋转3,座机4 | WLAN | 是1,否0 |
| 厚度 | 9~85 | 电池容量 | 100~4000 |
| 产品重量 | 48.4~790.2 | 重力感应器 | 是1,否0 |
| 屏幕数量 | 单1,双2 | 方向感应器 | 是1,否0 |
| 主屏幕尺寸 | 0~7 | 文字输入方法数 | 1~3 |
| 显示分辩率 | 6240~921600 | 智能系统 | 是1,否0 |

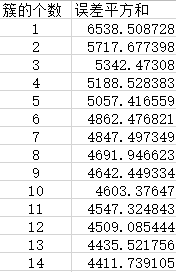
1. **聚类过程**

**（1）多次聚类尝试**

1. **只去掉属性id**

首先将数据项中的id属性去掉，每个产品都具有一个id，所以id是无效项，对于聚类来说没有什么实际意义。

之后将聚类结果sum of squared errors绘制了图表（其中横轴为簇的个数，纵轴为误差平方和）：

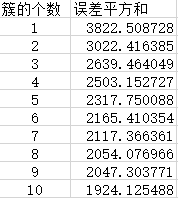


从图中可以看出误差平方和的范围在4400-6400之间，下降速率在1-6间下降较为快速，6之后下降变缓，说明当簇的个数超过7个之后，聚类的效果不再显著，虽然误差平方和还会继续下降，但是簇越多并不见得效果越好，如果每个产品一个类误差平方和肯定会变小，但是失去了聚类的意义。

1. **去掉属性id、型号、品牌**

考虑到品牌和型号很多，基本超过了30种，所以对于聚类为10个以下的类来讲，会出现把不同型号不同品牌的产品聚集在了一个类中，这样会影响聚类的效果，所以在第二次尝试中，我将型号和品牌这两个属性也去掉了。

聚类的结果的误差平方和如下：



从图中可以看出，误差平方和的范围在1900-3900之间，比上一次尝试（只去除id）的误差平方和小了一倍，效果十分显著。下降速率在1-6间下降较为快速，在6之后下降变缓。

1. **去掉其他属性值：**
2. 在去掉id、品牌、型号的基础上再去掉**上市时间（K=6、seed=10）：**



1. 在去掉id、品牌、型号的基础上再去掉**颜色（K=6、seed=10）：**



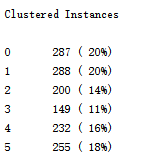
可以看出，和只去掉id、品牌、型号比较，去掉其他的属性并没有显著的变化，所以最终我们之去掉id、品牌、型号这三项。

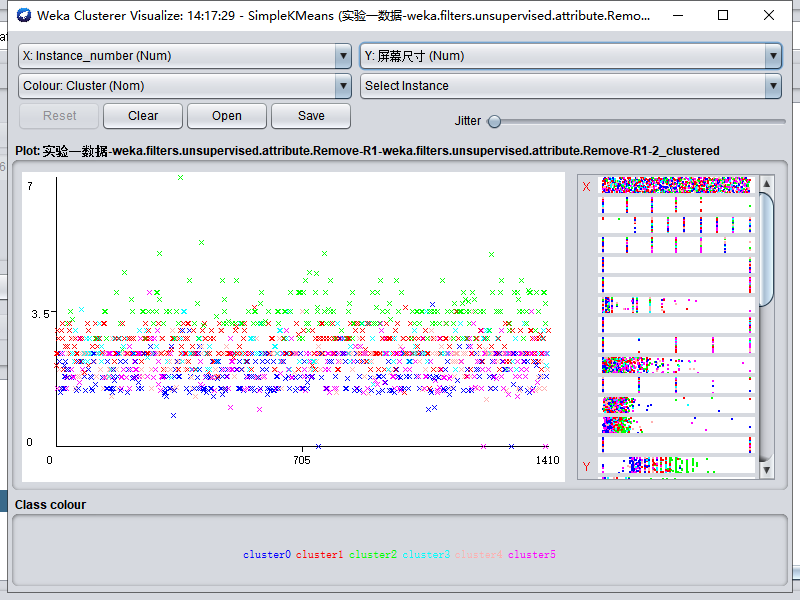
1. **确定k的值**

由上面的聚类实验可以得出，在簇的个数为6的时候，聚类效果的变化开始缓慢，因此我们将6作为最佳的簇数量，即k=6，并且采用第二次的数据处理（即去掉id、型号、品牌）。

从聚类结果中可以看出，本次聚类经过了19次迭代，且误差平方和约2165。

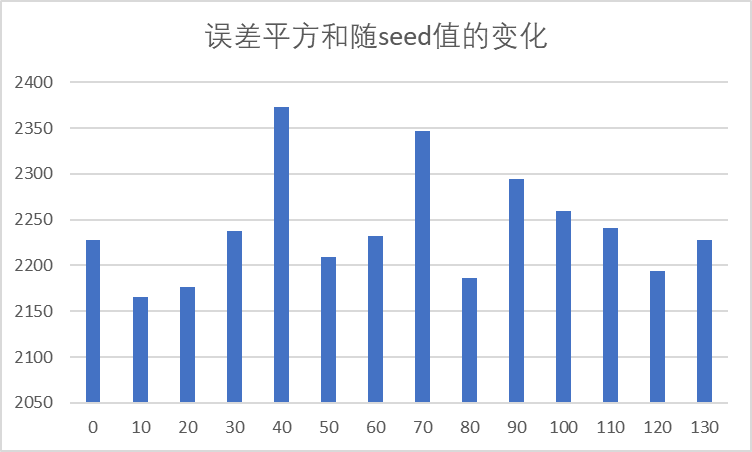


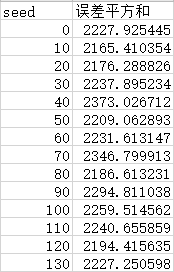




1. **改变初始簇心（调整seed的值）**

本次实验中，以10为间距从0开始来测试，测试的seed数有0、10、20、30、40、50、60、70、80、90、100、110、120、130

结果如图所示



从图中可以看出，seed=10的时候，误差平方和最小，为2165.41035402899

1. **模型分析**

经过以上的分析，最终我们选择的最佳模型为数据预处理去除id、型号、品牌这三项属性、k=6,seed=10的模型，误差平方和为2165.41035402899。