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

一、 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。

二、 实验内容

(a) 熟悉 WEKA 软件的使用；

(b) 给出本数据的数据描述报告；

(c) 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；

(d) 给出你认为最优的模型并加以解释。

三、 实验环境

Windows 环境、weka-3-8

四、 实验过程及结果

1. 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
显示分辨率	6240~921600	智能系统	是 1,否 0

2. 聚类过程

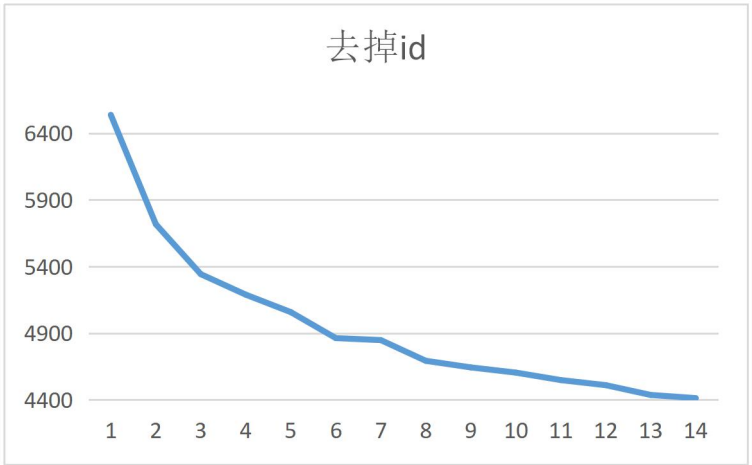
（1）多次聚类尝试

A. 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）：

簇的个数	误差平方和
1	6538.508728
2	5717.677398
3	5342.47308
4	5188.528383
5	5057.416559
6	4862.476821
7	4847.497349
8	4691.946623
9	4642.449334
10	4603.37647
11	4547.324843
12	4509.085444
13	4435.521756
14	4411.739105



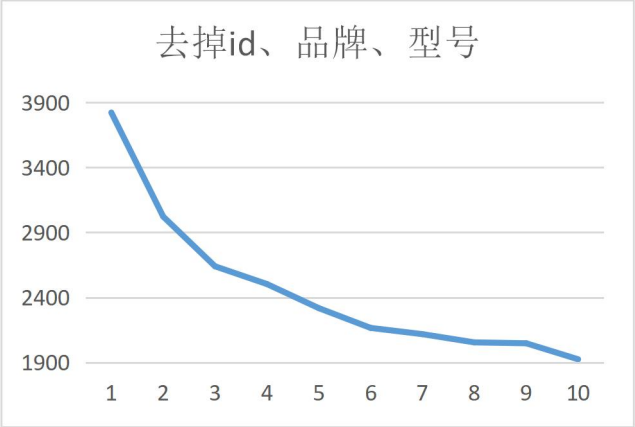
从图中可以看出误差平方和的范围在 4400-6400 之间，下降速率在 1-6 间下降较为快速，6 之后下降变缓，说明当簇的个数超过 7 个之后，聚类的效果不再显著，虽然误差平方和还会继续下降，但是簇越多并不见得效果越好，如果每个产品一个类误差平方和肯定会变小，但是失去了聚类的意义。

B. 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：

簇的个数	误差平方和
1	3822.508728
2	3022.416385
3	2639.464049
4	2503.152727
5	2317.750088
6	2165.410354
7	2117.366361
8	2054.076966
9	2047.303771
10	1924.125488



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。

C. 去掉其他属性值：

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

```
Number of iterations: 19
Within cluster sum of squared errors: 2128.615189521237
```

② 在去掉 id、品牌、型号的基础上再去掉**颜色（K=6、seed=10）：**

```
Number of iterations: 18
Within cluster sum of squared errors: 2133.9893365287717
```

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。

（2） 确定 k 的值

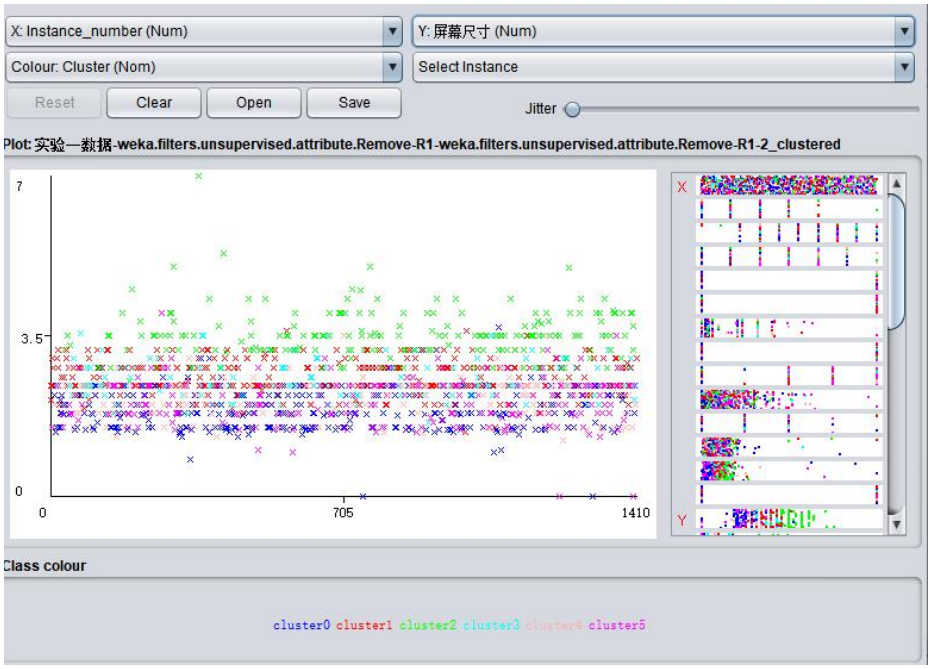
由上面的聚类实验可以得出，在簇的个数为 6 的时候，聚类效果的变化开始缓慢，因此我们将 6 作为最佳的簇数量，即 k=6，并且采用第二次的数据处理（即去掉 id、型号、品牌）。

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。

```
Number of iterations: 19
Within cluster sum of squared errors: 2165.410354028998

Clustered Instances

0       287 ( 20%)
1       288 ( 20%)
2       200 ( 14%)
3       149 ( 11%)
4       232 ( 16%)
5       255 ( 18%)
```

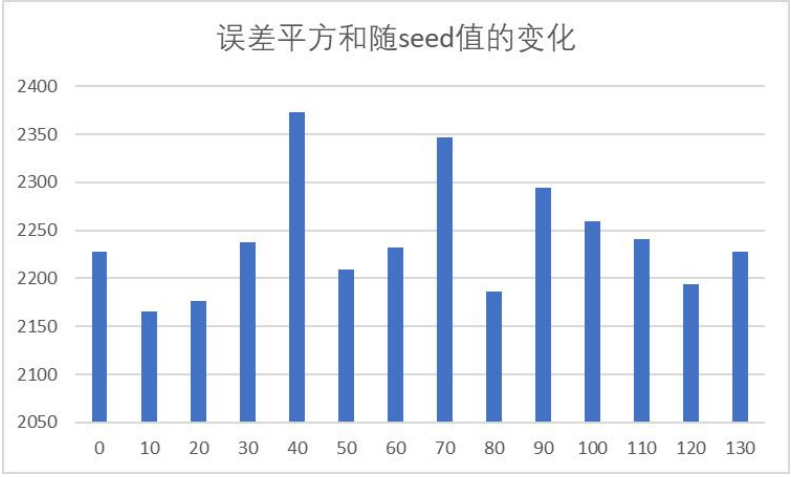


（3） 改变初始簇心（调整 seed 的值）

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

seed	误差平方和
0	2227.925445
10	2165.410354
20	2176.288826
30	2237.895234
40	2373.026712
50	2209.062893
60	2231.613147
70	2346.799913
80	2186.613231
90	2294.811038
100	2259.514562
110	2240.655859
120	2194.415635
130	2227.250598



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

3. 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。