北京邮电大学软件学院

2019-2020 学年第一学期实验报告

课程名称： 数据挖掘

项目名称： 实验一——聚类

项目完成人：

姓名： 罗舒婷 **学号：** 2017211951

指导教师： 牛琨

日　　期： 2019 年 10 月 27 日

一、 数据描述报告

宏观评价：

质量	完整性	准确性	合理性	一致性	时效性	可信性	可解释性
定义	主要包括实体缺失、属性缺失、记录缺失和字段值缺失四个方面；	一个数据值与设定为准确的值之间的一致程度，或与可接受程度之间的差异；	主要包括格式、类型、值域和业务规则的合理有效；	系统之间的数据差异和相互矛盾的一致性，业务指标统一定义，数据逻辑加工结果一致性；	当需要时，数据获得且是及时更新的；	数据是现实世界的真实反映	子空间内可解释该对象为什么和在何种程度和父空间内的呈现
评估维度	空值占比	异常值、箱线图、对比	合规记录的比率	合规率	时间划分	从业务逻辑角度判断	通过实验
评估	✔	✔	✔	✔	✔	✔	✔

各项属性数据：

属性数据质量	属性	离散/连续/标称
产品 ID		
型号		
品牌	标称	文字
颜色数	标称	1-7
上市时间	数值	离散
市场定位	标称	0-6
芯片平台	二元	
G 网	二元	
芯片主频	数值	连续
AP	二元	
频段数量	数值	离散
零售价格	数值	连续
外观类型	标称	0-4
厚度	数值	连续
产品重量	数值	连续
屏幕数量	二元	
主屏幕尺寸	数值	离散
显示分辨率	数值	离散
触摸屏	标称	0-2
键盘类型	标称	0-2
RAM	数值	离散
ROM	数值	离散

Flash 内存	数值	连续
摄像头	数值	离散
定位	二元	
FM 广播	二元	
电视	二元	
Modem	二元	
红外	二元	
蓝牙	二元	
WLAN	二元	
电池容量	数值	连续
重力感应器	二元	
方向感应器	二元	
文字输入方法数	标称	1-3
智能系统	二元	

数据给出了产品 ID、型号等三十六个属性，其中标称属性产品颜色分为 1-7，品牌属性为文字描述，市场定位用标称属性 0-6，表示不同定位，外观属性用 0-4，触摸屏 0-2，键盘类型 0-2，文字输入法为 1-3。外观类型。数值属性为 13 个，连续数值属性有 6 个，离散属性有 7 个。二元属性有 14 个，非对称二元属性有屏幕数量，芯片平台两项，是否布尔二元属性占 12 个。无缺失值，数据符合业务逻辑。

相关性：

1. 文字输入方法数和触摸屏、FM 广播、红外相关
2. 方向感应器和外观类型、ROM、电视、红外、WLAN、重力感应器相关
3. 重力感应器和芯片平台、蓝牙、WLAN、方向感应器和智能系统相关
4. 电池容量和型号、AP、屏幕数量、屏幕尺寸相关
5. WLAN 和分辨率、重力感应器、智能系统相关
6. 红外和品牌、市场定位、屏幕数量、电视、方向感应器、文字输入方法数相关
7. Modem 和型号、AP、屏幕尺寸、定位、电视相关
8. 电视和型号、频段数量、红外、方向感应器相关
9. FM 广播和型号、频段数量、厚度、蓝牙相关

10. 键盘类型和型号、屏幕数量、屏幕尺寸、触摸屏、定位、智能系统相关

11. 触摸屏和芯片主频、屏幕尺寸、键盘类型、WLAN、文字输入方法数、智能系统相关

12. 产品重量和零售价格、屏幕数量、屏幕尺寸相关

13. 厚度和上市时间、外观类型、RAM、FM 广播相关

14. 零售价格和品牌、颜色数、市场定位、芯片平台、频段数量、外观类型相关

15. 频段数量和 G 网、AP、零售价格、FM 广播、电视相关

16. G 网和型号、颜色数、频段数量、屏幕数量相关

17. 颜色数和品牌、上市时间、AP 零售价格、定位相关

二、 聚类

·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。

Relation: weka.filters.unsupervised.attribute.Remove-R1,9,23-24

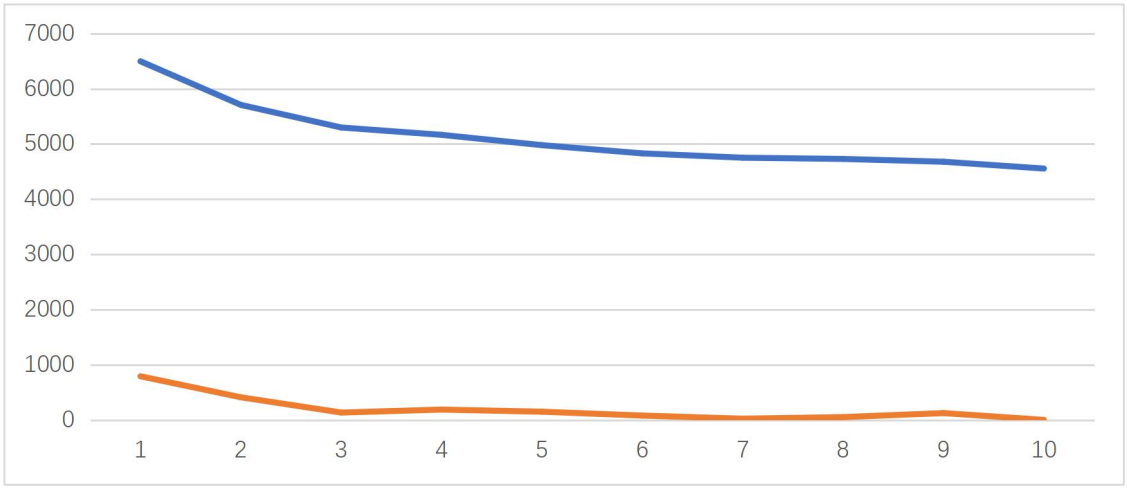
Instances: 1411

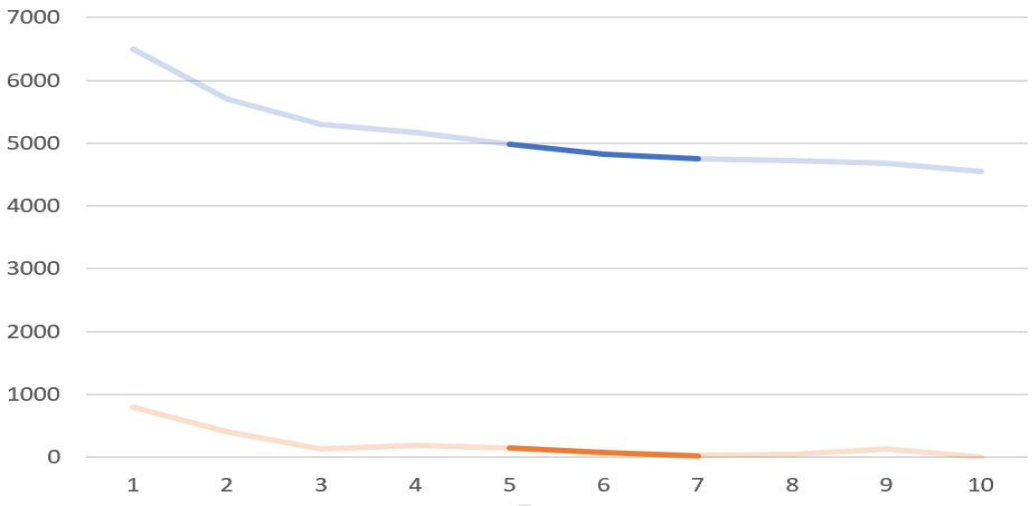
移除属性: 产品 ID，芯片主频，Flash 内存，摄像头

Attributes: 32

从图表可以看出，当属性值为 32 个，k 取值由 1-10 依次增加时，方差递减，其中 7-8 最为平滑。方差最小值出现在簇数 k=10，最大值出现在 k=1.

seed=10		
numCluster	Error	Error_i-Error_i+1
1	6498.34	790.036
2	5708.304	409.7327
3	5298.571	133.1306
4	5165.441	186.9361
5	4978.505	148.7428
6	4829.762	78.44843
7	4751.313	22.25048
8	4729.063	50.59465
9	4678.468	123.6174
10	4554.851	-





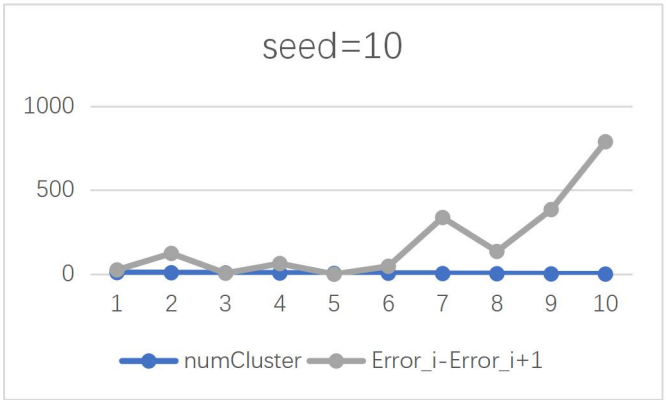
=== Run information ===

Relation: weka.filters.unsupervised.attribute.Remove-R1-3,9,23-24

Instances: 1411

移除属性: 产品 ID，型号，品牌 芯片主频，Flash 内存，摄像头

Attributes: 30



如图表所示，当属性为 30，seed=10，，k 由 1 至 10，方差递减，其中 5 与 6 簇方差相同，差值为 0。方差最小值出现在簇数 k=10，最大值出现在 k=1.

seed=10		
numCluster	Error	Error_i-Error_i+1
1	3782.34	788.1506168
2	2994.19	384.5611536
3	2609.628	135.6483029
4	2473.98	337.6243361
5	2136.356	0
6	2136.356	47.57061758
7	2088.785	63.64405788
8	2025.141	6.324722546
9	2018.816	123.9772419
10	1894.839	25.84124078

·改变初始簇心，研究簇心变化给聚类结果所带来的变化；

Relation: weka.filters.unsupervised.attribute.Remove-R1,9,23-24

Instances: 1411

移除属性: 产品 ID，芯片主频，Flash 内存，摄像头

Attributes: 32

numCluster: 7-8:

如图表所示，当属性为 32，簇的个数固定时，方差变化大小较无规律，方差浮动较为平缓，其中当簇的个数为 7 时，seed 个数 7 与 8 时方差差距最小，3 与 4 之间的差值为第二小的差值，最小的方差值为簇心 seed 数量为 7 的时候，最小值 4724.244，其次小为簇数是 6 的时候。最大值出现在 seed 数为 8 时，值为 4813.764.

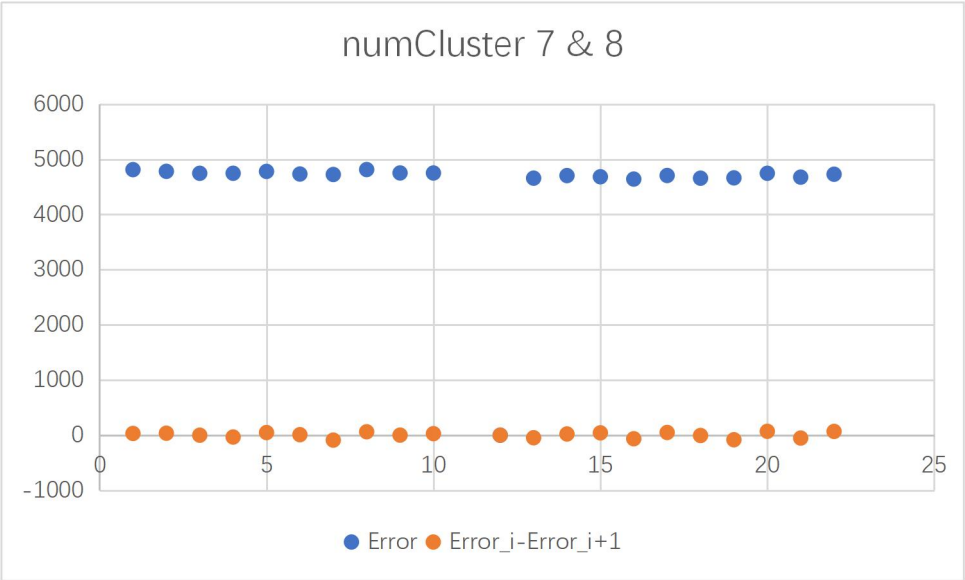
当簇的个数固定为 8 时，最大值出现在 seed 数为 8 时，值为 4746.251，最小值出现在 seed 为 4 的时候、其次小是 seed 为 6 的时候. 3 与 4 之间的方差差值为最小的差值即 3 与 4 较为平缓，

两组的共同点为，方差值最大出现在 seed=8，seed=6 时，两组数据的方差值

都较小。当 seed=3 与 4 时，差值变化线较为平缓，方差变化小。

numCluster=7		
seed	Error	Error_i-Error_i+1
1	4812.658	31.06998228
2	4781.588	35.55354822
3	4746.034	-0.38831794
4	4746.422	-33.63006271
5	4780.053	47.18220477
6	4732.87	8.646672653
7	4724.224	-89.54047968
8	4813.764	61.45886188
9	4752.305	0.991779182

numCluster=8		
seed	Error	Error_i-Error_i+1
1	4658.053	-46.57441428
2	4704.628	21.45792989
3	4683.17	41.86661326
4	4641.303	-64.8424043
5	4706.145	48.82648966
6	4657.319	-6.051835007
7	4663.371	-82.88065943
8	4746.251	69.80546594
9	4676.446	-52.61702976
10	4729.063	67.98946567



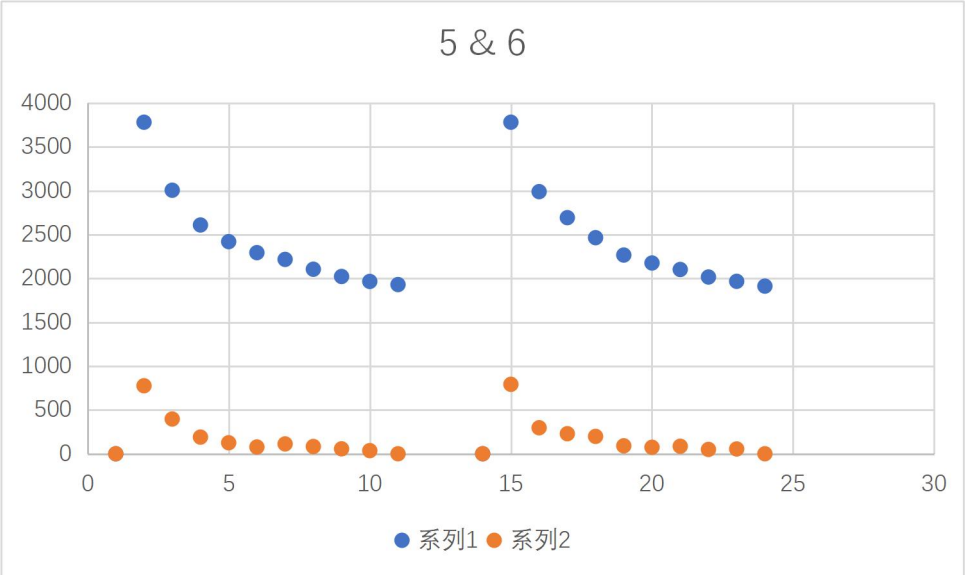
Relation: weka.filters.unsupervised.attribute.Remove-R1-3,9,23-24
Instances: 1411
移除属性: 产品 ID，型号，品牌 芯片主频，Flash 内存，摄像头
Attributes: 30
numCluster: 5-6

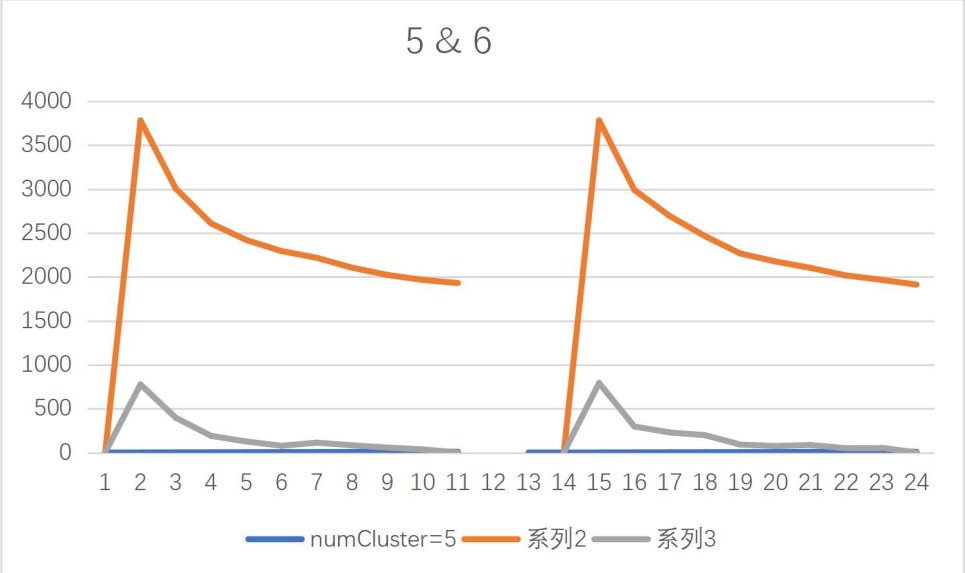
如图表所示，属性为 30 个时，当簇的个数固定，方差变化大小较无规律，方差浮动呈现递减趋势，其中当簇的个数为 6 时，seed 个数 5 与 6 时方差相同，差值为 0。此为最小差值。

当簇数为 7 时

numCluster=5		
seed	Error	Error_i-Error_i+1
1	3782.34	775.943811
2	3006.396	396.4780176
3	2609.918	190.0423429
4	2419.876	125.5143307
5	2294.362	77.04208909
6	2217.32	112.1929016
7	2105.127	82.73931857
8	2022.387	56.39903862
9	1965.988	35.83047017
10	1930.158	-

numCluster=6		
seed	Error	Error_i-Error_i+1
1	3782.34	791.9733913
2	2990.367	296.6994764
3	2693.667	228.5751189
4	2465.092	198.1024015
5	2266.99	90.73804024
6	2176.252	74.04000785
7	2102.212	85.68888207
8	2016.523	49.29446339
9	1967.228	54.85152753
10	1912.377	-





三、选择最佳模型说明理由：

我选择属性为颜色数，上市时间，市场定位，芯片平台，G 网，AP，频段数量，零售价格，外观类型，厚度， 产品重量，屏幕数量，屏幕尺寸，分辨率，触摸屏，键盘类型，RAM， ROM，定位，FM 广播，电视，Modem，红外，蓝牙， WLAN，电池容量，重力感应器， 方向感应器，文字输入方法数，智能系统 K=6，seed=10。理由：K 均值是一种简便、实用、无监督聚类分析算法。这种算法在已知簇数时，可以很好地实现数据的聚类分析。随机选择 k 个数据点做聚类中心，计算其他点到这些聚类中心点的距离，通过对簇中距离平均值 d 饿计算，不断改变这些聚类中心的位置，优势是算法简单，执行和收敛过程相对较快，是一种常见的聚类算法，K=6，seed=10 时，方差值最小。