基于 k-means 聚类的手机受欢迎程度研究

席讴婕

（内蒙古电子信息职业技术学院，内蒙古 呼和浩特 010011）

摘要：本文以京东商城为例，利用 Python 语言设计并实现了京东商城的手机销售信息爬取，提取手机销售信息的四个特征，然后用 k-means 聚类算法将这些手机分成三类，分为“最受欢迎的手机”、“较受欢迎手机”、“普通手机”，构建出手机受欢迎度聚类模型，为商家生产、销售手机提供依据。

关键词：网络爬虫 Python AJAX Selenium K-means

随着网络的发展，网上购物的人日益增多，对于商家来说，网上销售渠道和传统商业实体一样占有举足轻重的地位，如何提升网上销售渠道的销量，成为商家关注的首要问题，针对这个问题，根据“运行内存”、“ 存储内存”、“ 价格”和“销量”四个基本特征，采用 k-means 聚类算法对手机进行聚类分析，构建手机受欢迎程度聚类模型，为解决上述问题提供依据。

1 相关技术

1.1 网络爬虫

网络爬虫是一种高效的信息采集利器，它可以依照一定的规则，自动采集网页内容。网络爬虫主要分通用爬虫和聚焦爬虫两种。

通用爬虫，将网页完整的下载到本地，由于没有爬取范围限制，爬虫会不停地爬取直到抓完所有数据，形成互联网内容的完整备份，一般用于搜索引擎抓取系统，如百度、谷歌、搜狗等。

聚焦爬虫，面向特定需求的一类网络爬虫，可以对诸如行业、内容、发布时间、页面大小等很多因素进行细致筛选，尽量保证只抓取与需求相关的网页信息。

1.2 Python

Python 是一个高层次的解释型、面向对象、动态数据类型的脚本语言。是数据分析师的首选数据分析语言，也是硬件的首选语言，语法优美、代码简洁、开发效率高、支持的模块多，相关的 HTTP 请求模块和 HTML 解析模块非常丰富。

1.3 AJAX

AJAX 是 Asynchronous JavaScript and XML 的缩写，可以在不重载整个页面的情况下对网页的某部分进行更新，可以缩短用户响应时间，减少不必要的数据传输、时间及降低网络上数据流量。

1.4 Selenium

Selenium 是一个 Web 的自动化测试工具，可以模拟浏览器的行为，能获取到浏览器能请求到的所有数据，主要用于动态渲染页面的爬取。浏览器能请求到的，使用 selenium 也能请求到，而且更稳定，缺点是代码量多，性能低。

1.5 K-means 聚类算法

K-means 聚类算法属于无监督的机器学习，聚类无需人工干预，具体的过程如下：初始时从 n 个样本中随机生成 k 个中心点，计算其它（n-k）个数据点到这 k 个中心点的距离，每个中心点代表一个簇，将（n-k）个中心点归于距离最近的簇中，计算每个簇的平均值作为簇的新的中心点的值，重复上述的迭代过程到准则函数收敛，没有收据移动。最终把数据划分成不同数据集合，使得评价聚类性能的平方误差准则最优，即不同类间相似性较低，同类间相似性较高。

2 爬取京东产品信息

2.1 特征数据采集

本文数据来源于京东网上商城，利用爬虫于 2019 年 4 月 15 日从京东网页上爬取 60 种手机信息，每种提取四个特征，共得到 240 条数据，分别是运行内存、存储内存、价格和销量，得到一个 60*4 特征矩阵。为用户的聚类分析做好准备。

表 2.1 初始特征值

运行内存(GB)	存储内存(GB)	价格(元)	销量（万台）
8	128	5999	3
0	128	5699	102
8	128	3298	10
4	64	1299	148
8	64	3988	148
4	64	1199	63
6	128	1599	10
8	128	3299	15
6	64	1499	0
4	32	899	44

2.2 原始数据预处理

数据操作会出现噪音数据、数据不完整、数据不一致的情况，噪音数据是指错误值，数据不完整指有些数据出现属性值空缺，数据不一致是指数据的数量级不同。本文中提取的特征值主要有数据不一致现象，其中运行内存在 10 数量级，存储内存在 1000 数量级，价格在 10000 数量级，K-means 算法基于距离进行分类，数量级的显著差异对整体分类的精度有很大的影响，所以需要进行预处理，对运行内存乘以 0.01，存储内存乘以 0.001，价格乘以 0.0001，这些数据落到[0.0,1.0]这个区间，这样会提高整体分类准确率。

2.3 使用 K-means 方法聚类，聚类后结果如下：

本文使用 K-means 算法对数据正规化数据分类，比较后得到分成 3 类时聚类效果最好，即此时任意两个待分类对象的欧式距离和最小。

2.4 结果分析及小结

根据 K-means 算法将上述数据分为 3 类，分别为“最受欢迎的手机”、“较受欢迎手机”、“普通手机”，其中第二类平均销量最高，为“最受欢迎的手机”，共有九个样本，其中五个样本品牌为苹果，二个样本品牌为三星，二个样本品牌为华为。平均运行内存为 3GB，平均存储内存为 96GB，平均价格为 6400 元，占样本总数的 15%，第三类手机平均销量次之，为“较受欢迎手机”，平均运行内存为 6GB，平均存储内存为 125GB，平均价格为 3120 元，占样本总数的 37%，第一类数据平均销量最低，为“普通手机”，平均运行内存为 4.7GB，平均存储内存为 59GB，平均价格为 1310 元。

3 结束语

从以上的分析可以看出，“最受欢迎的手机”市场占比最小，仅为 15%，而销量却在平均 65 万台，而“普通手机”市场占比为 48%，几乎占领一半的市场，平均销量却是最低，平均销量为 32 万台，可以得出结论，一味的追求低价、高运行内存和高存储内存并不能使销量升高，打造我们的优质的国产品牌，才能在手机销售竞争中立于不败之地。

参考文献：

[1]Liu B,Chang K C C.Sepecial issue on web content mining[J].ACM SigKddExplorations,2004,6(2):14

[2]Anil K J.Data clustering:50 years beyond K-Means[J].Pattern Recognition Letters,2010,31(8):651-666