

【文章编号】 2096-2835(2020)04-0482-08 DOI:10.3969/j.issn.2096-2835.2020.04.011

基于改进 K -means 算法的电子 商务客户细分研究

靖立峥, 吴增源

(中国计量大学 经济与管理学院, 浙江 杭州 310018)

【摘要】 目的: 针对传统 K -means 算法需要人为设定 K 值的缺陷, 提出改进的 K -means 算法, 并将其应用于电子商务客户细分研究。方法: 首先, 在经典 RFM 模型的基础上, 增加客户消费行为特征; 其次, 为确定最佳聚类数目, 引入 CH 评价指标, 以对 K -means 算法进行改进; 最后, 选取了包含 37 376 个样本的电子商务客户数据集进行实证研究。结果: 与拐点法相比, 通过 CH 指标确定 K 值更加直观; 与谱聚类相比, 加入 CH 指标的 K -means 算法具有更优的聚类效果及运行效率。结论: 结合 CH 聚类评价质量指标和 K -means 算法能有效提高电子商务客户细分的准确性和效率。

【关键词】 K -means 聚类; 客户细分; 消费行为偏好

【中图分类号】 TP391; F274

【文献标志码】 A

Research on e-commerce customer segmentation based on improved K -means clustering

JING Lizheng, WU Zengyuan

(College of Economics and Management, China Jiliang University, Hangzhou 310018, China)

[Abstract] **Aims:** In view of the shortcoming of setting K value artificially in the traditional K -means algorithm, we proposed an improved K -means algorithm. This algorithm was employed to segment e-commerce customers. **Methods:** Firstly, on the base of the RFM model, the features of customer consumption behavior was introduced. Secondly, in order to determine the optimal number of clustering, the CH index was introduced to improve the K -means algorithm. Finally, a dataset with 37,376 samples of e-commerce customers was selected for empirical research. **Results:** Compared with the inflection point method, K values selected by the CH index was more intuitive. Compared with the spectral clustering, the clustering effect and operating efficiency of the K -means algorithm with the CH index were better. **Conclusions:** The accuracy and efficiency of e-commerce customer segmentation could be improved effectively by combining the CH index and the K -means algorithm.

[Key words] K -means clustering; customer segmentation; consumer behavior preference

【收稿日期】 2020-09-05

《中国计量大学学报》网址: zgjl.cbpt.cnki.net

【基金项目】 国家自然科学基金项目 (No. 71871205), 浙江省自然科学基金项目 (No. LY20G010008)

【通信作者】 吴增源 (1979-), 男, 副教授, 博士, 主要研究方向为信息管理、机器学习。E-mail: wuzengyuan@cjl.u.edu.cn

凭借方便快捷、低成本、不受时空限制等优势,全球电子商务蓬勃发展。今年上半年,在新冠疫情防控期间,电子商务在保障社会基本生产、生活物资供应和拉动经济增长方面起到了重要作用。良好的客户关系管理是电商平台构建竞争优势的重要一环^[1-2]。百雀羚广告《一九三一》的低转化率以及频频爆出的直播带货失败案例表明,高效的客户关系管理必须以客户细分为基础^[3-4]。客户细分是指利用统计分析或数据挖掘等技术分析客户数据,对客户进行分类,以提供有针对性、个性化的服务,从而提高营销效率^[5]。只有深入了解不同客户群体的偏好和需求,才能实施精准营销。

在电子商务客户细分研究领域,基于机器学习的聚类分析技术是目前较常使用的方法之一。该方法以经典的 RFM 模型^[5]为基础,利用多维度指标,精准体现不同客户群体的消费行为和习惯。但是,现有研究侧重于使用客户的历史订单数据,无法全面反应不同客户群体的行为偏好和消费习惯。另外,经典的 K-means 算法存在人为设定 K 值的缺陷。因此,在特征选择过程中,本文增加了客户消费行为数据:客户消费间隔(recency, R)、消费次数(frequency, F)、浏览(details page of product, P)、加购(add to cart, A),以及收藏次数(favourite, V);在算法改进方面,引入 Calinski-Harabasz(CH)聚类质量评价指标,优化 K 值选取。实验结果表明:使用 CH 聚类评价质量指标优化传统 K-means 算法,能有效提高电子商务客户细分的效率和准确性。

1 文献回顾

很多学者从不同的角度提出了客户细分方法^[6],这些方法依据选取的客户指标的不同主要分为三类:基于人口统计的视角、基于客户生命周期的视角和基于客户行为的视角。

基于人口统计^[7]视角的客户细分方法主要采用问卷调查的手段,以客户年龄、性别、家庭收入、婚姻状况教育程度等为基础划分成不同的群体。但是这种单纯地利用人口统计指标细分客户的方法不能预测客户的未来购买行为^[8],也不能对客户流失的原因做出精准的分析,企业难以采取具体的针对性措施去吸引客户、保持客户。

基于客户生命周期视角^[9]的客户细分方法的具体过程是:企业利用新增、留存、流失的客户数量信息,绘制客户生命周期分布曲线。根据客户所处的生命周期的不同阶段,对其采取不同的措施。客户忠诚度分类方法^[10-11]是现有基于客户生命周期的细分研究中最常用的细分方法。Christopher & Payne(2008)根据客户所处的生命周期的不同阶段分为潜在顾客、现实买主、长期客户、支持者和鼓吹者^[12]。但是这种细分方法无法体现客户价值特征之间的差异,以及客户的偏好特征和行为特征的差异。

近年来,随着数据挖掘技术的不断发展,基于客户行为的客户细分方法逐渐成为研究热点。这类细分方法主要采用的是基于机器学习的聚类分析技术,利用多维度指标,精准体现不同客户群体的消费行为和习惯^[13]。在这种方法中,作为经典的客户价值模型-RFM 模型已经成功应用于不同行业的客户细分^[14-15]。但是,RFM 模型中的 F、M 指标会存在多重共线性的问题^[16]。由于以上模型的缺陷,不同行业的研究学者根据行业特点对其进行了改进和扩展^[17-19]。Serhat Peker(2017)针对杂货店的客户,利用 K-means 聚类算法,提出了基于 LRFMP 的客户细分模型,但是仍没有解决 F 和 M 之间的多重共线性的问题^[13]。任春华等(2019),针对汽车销售行业的特点,提出了汽车忠诚客户的 LRFAT 模型。该模型可以消除 F 值与 M 值的共线性问题,并且能识别出忠诚客户^[16]。

虽然引入客户行为特征数据之后,电子商务客户细分研究取得了一定的研究成果,但是仍有以下两个问题没有得到解决:一是对于客户细分模型指标的选取,现有研究侧重于使用客户的历史订单数据,但是忽略了客户的线上行为数据,无法更加全面的反映不同客户群客户的行为偏好和消费习惯;二是经典的 K-means 算法虽然原理简单,易于实现,但存在人为设定 K 值的缺陷。由于每个行业客户的结构分布不同,人为设定 K 值,很有可能造成细分结果与实际有较大差异。因此,在特征选择上,本文构建了融入客户线上行为特征的 RFPV 客户细分模型;在算法上,本文借助 K-means 算法进行聚类分析的基础上,引入 CH 聚类质量评价指标,以期实现更优的聚类效果。

2 基于改进 K-means 的电子商务客户细分方法

2.1 客户细分指标选取

RFM 模型最早由 Hughes(1994)^[20] 提出,一般用来衡量客户的价值特征,识别高价值客户。自提出以来,RFM 模型已经广泛应用于客户细分,R(recency)通常代表客户最近一次消费据观察期截止日期的消费间隔,R 值越大代表客户据上一次消费的日时间越长,客户价值越低;F(frequency)通常代表观察期内的消费的次数,F 值越大代表客户消费越频繁,客户价值越高;M(monetary)通常代表客户在观察期内消费的总金额,M 值越大,客户价值越高。

虽然传统 RFM 模型已广泛应用于各个行业的客户细分问题,但是仍存在两个问题:第一,RFM 模型中的 F 和 M 指标存在多重共线的问题;第二,使用 RFM 模型进行客户细分,不能体现客户在该电商平台的活跃度,以及不同客户群之间消费和行为习惯的差异。随着大数据技术的提升,从电商平台提取的客户信息数据的维度增多,这些数据更加细致全面的反映了客户的价值特征、消费习惯以及行为偏好,因此,基于传统 RFM 模型,本文融入了客户的线上行为指标,提出了针对电子商务客户细分的 RFPV 模型,其中,P、A、V 指标可以体现客户的活跃度、线上消费习惯。

2.2 基于优化 K 值选取的改进 K-means 算法

2.2.1 K-means 算法描述

K-means^[21] 是一种经典的基于划分的聚类方法,一般用欧氏距离作为衡量两个数据点之间相似度的指标,相似度越大,距离越小。该算法的核心思想是:首先确定聚类数目 K 和 K 个初始聚类中心。根据数据点与聚类中心的距离,不断更新聚类中心的位置,使得每个簇的误差平方和(sum of squared error, SSE)变小。当 SSE 不再变化或者目标函数收敛时,SSE 值达到最小,迭代停止,得到最终的聚类结果。其算法流程如下。

1) 初始化聚类中心 确定聚类数目 K,从数据集中随机选取 K 个点作为初始聚类中心 C_i ($1 \leq i \leq K$)。

2) 分配样本 计算其余数据点与聚类中心

C_i 的欧氏距离,找出最短距离并将所有样本分配到聚类中心 C_i 对应的簇中。

欧式距离计算公式为

$$d(x, C_i) = \sqrt{\sum_{j=1}^m (x_j - C_{ij})^2} \quad (1)$$

式(1)中, x 为数据对象, C_i 为第 i 个聚类中心, m 为数据对象的维度, x_j 、 C_{ij} 为 x 和 C_i 的第 j 个属性值。

3) 更新聚类中心 计算每一簇中所有点的平均值及平方误差,将平均值作为新的聚类中心,重复步骤 2)。

平方误差计算公式为

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} |d(x, C_i)|^2 \quad (2)$$

4) 直至聚类中心不再变化或达到最大的迭代次数,循环结束,得到最终聚类结果。

2.2.2 优化 K 值选取的改进 K-means 算法

尽管 K-means 聚类算法不依赖于数据的标签信息,对大数据特征挖掘有很好的解释效果,但是传统的 K-means 算法存在人为设定 K 值的缺陷。基于此,本文引入了 Calinski-Harabasz(CH) 聚类质量评价指标,将最高的 CH 得分所对应的类设定为本研究的聚类数目。

CH 指标由簇间样本分离度与簇内样本紧密度的比值得到,CH 越大代表类自身越紧密,类与类之间越分散,即聚类结果更优。当簇内密集且簇间分离较好时,从 CH 得分折线图中可以明确得出最佳聚类数目,并且具有计算速度快的优点。

CH 指标优化 K-means 算法的实现步骤如下:

输入:数据集 $X = \{x_1, x_2, \dots, x_n\}$ (n 表示数据点的数量)。

1) 确定最佳聚类数目 k 。

① 计算 WGSS

WGSS(Within-Groups Sum of Squared Error)为簇内平方误差和,用来度量簇内样本的紧密度, WGSS 越小,簇内越紧密,聚类效果越好。其计算公式为

$$WGSS = \frac{1}{2} [(n_1 - 1) \bar{d}_1^2 + \dots + (n_k - 1) \bar{d}_k^2] \quad (3)$$

式(3)中, \bar{d}_k^2 为第 k 簇内样本的平均距离, n_k 为第 k 簇的样本数量。

② 计算 BGSS

BGSS(between-groups sum of squared error)为簇间平方误差和,用来度量簇间样本的分

离度, BGSS 越大, 簇间越分散, 聚类效果越好。其计算公式为

$$BGSS = \frac{1}{2}[(k-1)\bar{d}^2 + \sum_{i=1}^k (n_i - 1)(\bar{d}_i^2 - \bar{d}^2)] \quad (4)$$

式(4)中, \bar{d}^2 为所有样本间的平均距离, \bar{d}_i^2 为第 i 簇簇内样本的平均距离; n_i 为第 i 簇的样本数量; k 为样本集簇数。

③计算 CH 值

CH 得分值 S 的数学计算公式为

$$S(k) = \frac{\text{簇间平方误差和}}{k-1} / \frac{\text{簇内平方误差和}}{n-k} = \frac{BGSS}{WGSS} \times \frac{n-k}{k-1} \quad (5)$$

WGSS 越小, BGSS 越大, 则 CH 指标值越大, 聚类效果越好。

④绘制 CH 折线图, 确定最佳聚类数目 k 。

2) 运行 K-means, 得出聚类结果:

①初始化聚类中心

输入聚类数目 k , 从数据集中随机选取 k 个点作为初始聚类中心 $C_i (1 \leq i \leq K)$ 。

②分配样本

计算其余数据点与聚类中心 C_i 的欧氏距离, 找出最短距离并将所有样本分配到聚类中心 C_i 对应的簇中。

③更新聚类中心

计算每一簇中所有点的平均值及平方误差。将平均值作为新的聚类中心, 重复步骤 2)。

④直至聚类中心不再变化或达到最大的迭代次数, 循环结束, 得到最终聚类结果。

输出: 聚类结果 $C = \{c_1, c_2, \dots, c_k\}$ 。

2.2.3 对比验证

为了验证本文采取方法的有效性, 本文进行了两次对比实验: 第一次为聚类算法对比; 第二次为聚类质量评价指标的对比。

1) 算法对比

为验证本文算法的有效性, 选取 UCI 数据库中两个标准的测试数据集 Breast cancer 和 Iris plants 进行实验, 并采用相关研究中常用的谱聚类(Spectral Clustering)方法, 与本文的 K-means 算法进行对比研究。在算法评价上, 主要比较了两种算法在两个数据集上的聚类准确率以及运行时间, 如表 1。

表 1 两种算法在不同数据集上的运行情况

Table 1 Performance of the two algorithms working on different datasets

聚类算法	数据集			
	Breast cancer		Iris plants	
	准确率/%	时间/ms	准确率/%	时间/ms
K-means	0.854	19.8	0.893	21.9
谱聚类	0.667	103.8	0.900	118.1

从表 1 可以看出: 第一, 在聚类准确性上, K-means 在两个数据集上的准确率分别为 85.4% 和 89.3%, 都有比较好的聚类效果, 谱聚类的准确率分别为 66.7% 和 90%, 聚类效果相较 K-means 更不稳定; 第二, 在运行时间上, K-means 在两个数据集上的运行时间分别为 19.8 ms 和 21.9 ms, 而谱聚类的运行时间分别为 103.8 ms 和 118.1 ms, 可以明显看出, 谱聚类的运行时间是 K-means 算法的 5 倍多, 聚类效率较低。这主要由于谱聚类算法涉及数据之间的相似度矩阵, 在数据集维度较高时, 无法表现出更优的准确性和聚类效率。然而, 电商平台交易量大, 数据规模庞大且维度复杂, 对聚类效率的要求高, 需要及时根据客户的消费行为数据进行划分, 以实现高效精准的客户管理。因此, 对于电子商务平台的客户细分, 相比谱聚类, K-means 的聚类效果更优。

2) 聚类质量评价指标对比

针对传统 K-means 算法存在人为设定 K 值的缺陷, 本文引入 CH 指标, 作为客户细分时决定 K 值的依据。为了验证 CH 指标对于电子商务行业中的客户细分的适用性, 本文使用实证部分的电商数据集, 选取拐点法与 CH 进行对比。图 1 和图 2 分别为 CH 指标和拐点法的实验结果。

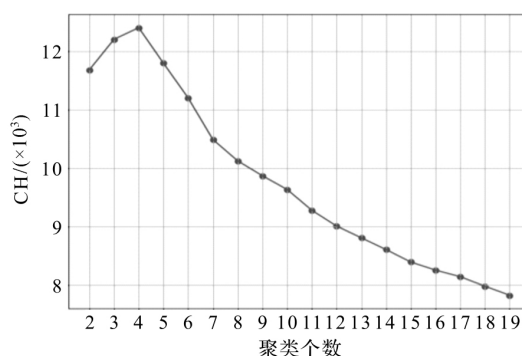


图 1 CH 得分折线图

Figure 1 Line chart of CH score

由图 1 可以看出,CH 得分折线图呈现先上升后下降的趋势,当聚类数目为 4 时,CH 值最高。因此,使用 CH 指标可以明确得出:该电子商务平台数据集的最佳聚类数目为 4。

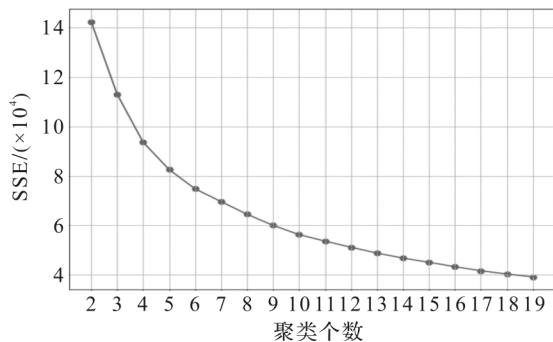


图 2 拐点法折线图

Figure 2 Line chart of inflexion point method

在图 2 中,横轴为聚类个数的变化,纵轴为数据的 SSE 方差,即凝聚度。拐点法原理是:在折线图的拐点处取得最佳聚类数目。其原因在于:拐点之后继续增加 K 值,对分类的准确度增加不高,但会增加簇数。从而导致对数据的划分过细,影响聚类效果。但是,从图 2 中可以看出,当 K 值从 4 变化到 6 时,折线图的变化较为平滑,即没有明显的拐点,无法准确判断出最佳聚类个数。

因此,与图 1 中 CH 得分的变化趋势相比,拐点法不适合用于电子商务客户细分的研究,使用 CH 指标优化 K -means 算法对于电子商务客户细分更具有有效性。

3 电子商务客户细分实证研究

3.1 数据集描述

本文研究数据是从国内某电子商务平台中提取的客户消费数据,数据集包含了 2014 年 11 月 18 日至 2014 年 12 月 18 日之间,一个月内的该平台移动端的 37 376 名客户的订单及线上行为数据。该数据集可用于以下实际问题的研究:1)建立客户流失预测模型,预测客户流失;2)研究用户在不同时间尺度下的行为规律,找到用户在不同时间周期下的活跃规律;3)根据数据集中的客户订单信息及行为和消费偏好信息,找出最具付费价值的客户群,对这部分客户行为进行分析;4)将数据集中的核心信息整合成几个指标,进行客户细分研究,针

对不同特征的客户群采取针对性措施。表 2 为本文研究涉及数据集中的字段及字段描述。

表 2 数据集中涉及的字段及字段描述

Table 2 Fields and descriptions involved in the dataset

字段名	数据类型	字段描述
User ID	Int	用户标识
Item ID	Int	商品标识
Category ID	Int	商品分类标识
Behavior type	String	行为类型,包括浏览、收藏等
Timestamp	Int	行为时间

3.2 数据处理

3.2.1 数据清洗

2014 年 11 月 18 日至 2014 年 12 月 18 日之间该电子商务客户的行为数据大约一亿多条,需要进行数据清洗:首先处理缺失值和异常值的数据,如购买费用为 0 的数据、购买日期为空以及购买费用明显有误的数据;然后处理重复数据。用户的购买行为由于时间精确到小时,会存在少量用户在一小时内重复购买或浏览统一商品的行为,因此不对此部分数据进行处理;最后,处理数据的一致性问题,本文涉及的指标 L、R 涉及到时间特征,并且以天为单位。而时间数据中的日期和小时存在于一列中,故将其拆分成两列。另外,本文将 Timestamp 字段中的时间戳字段类型转换成年月日的形式,以方便计算时间。

3.2.2 指标提取及归一化

结合本文提出的 RFPAV 模型,计算整理数据得到 37 376 位客户的数据,表 3 为部分数据。

表 3 RFPAV 模型部分数据表

Table 3 Partial data of RFPAV model

User ID	R	F	P	A	V
1	0	0	54	0	0
13	0	0	7	4	0
19	0	0	74	1	0
...
1 017 997	6	2	90	0	3
1 018 000	0	0	181	4	0
1 018 011	2	1	41	0	0

为了防止各指标单位不同所引起的差距过大,在进行算法实验之前,需要对指标提取之后的数据集进行归一化处理。本文采用的归一化方法为 Z-score 标准化方法,这种方法给予原始数据

的均值 (mean) 和标准差 (standard deviation) 进行数据的标准化。经过处理的数据符合标准正态分布,即均值为 0,标准差为 1,转化函数为

$$X^* = \frac{X - \mu}{\sigma}。$$

(6)

式(6)中, μ 为所有样本数据的均值, σ 为所有样本数据的标准差。

表 4 为归一化处理前后的部分数据,指标 R、F、P、A、V 的数据全部被转换为无量纲的数据值。

表 4 RFP AV 模型归一化处理部分数表

Table 4 Partial data of RFP AV model after normalized treatment

User ID	R	F	P	A	V
1	-1.249 537	-0.713 743	-0.469 836	-0.745 080	-0.390 554
13	-1.249 537	-0.713 743	-1.052 897	-0.223 663	-0.390 554
19	-1.249 537	-0.713 743	-0.221 725	-0.614 726	-0.390 554
...
1 017 997	0.623 415	-0.018 191	-0.023 237	-0.745 080	0.003 472
1 018 000	-1.249 537	-0.713 743	1.105 669	-0.223 663	-0.390 554
1 018 011	-0.625 219	-0.365 967	-0.631 109	-0.745 080	-0.390 554

3.3 实证结果分析

根据本文 2.2.3 小节中的实验结果,确定最佳聚类数目 $K=4$ 。通过 K-means 算法和 RF-PAV 模型,得到 4 组客户聚类数据,图 3 为 4 组客户分布饼状图,图 4 为 4 组客户 RFP AV 模型各指标分布图。

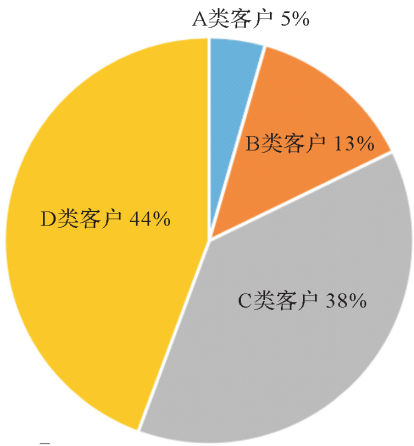


图 3 客户类型分布图

Figure 3 Distribution chart of customer types

从图 3 和图 4 可以看出,价值最大的客户群体是 B 类客户,包括 4 984 名顾客,占该电子商务平台数据集的 13%。与其他三种客户类型的指标值相比:B 类客户的 R 值较小,说明近期有在该电商平台进行消费;F 值最高,说明订单频率高,属于该电商平台中的活跃客户;P 值和 A 值

均最高,说明这类客户经常在该电商平台浏览商品,将商品加入购物车的频率也高。这可看出这类客户群当前价值和增值潜力最大,可将其归类为该电商平台的高价值客户群;V 值较低,说明这类客户群在遇到感兴趣的商品时,习惯于加购而不是收藏。针对这类客户群,企业需要投入主要精力和资源,保持和发展与这些客户的关系,充分挖掘这类客户的消费潜力。

第二类较有价值的客户群体是 A 类客户,包括 1 661 名客户,占该电子商务平台数据集的 5%。与其他三种客户类型的指标值相比:A 类客户的 R 值相较 B 类更小,说明在最近几天有在该电商平台进行消费。但是该类客户群的 F 值更低,说明消费频率比 B 类低,属于该电商平台中的较活跃客户;P 值较高,说明这类客户也经常在该电商平台浏览商品;与 B 类不同的是,A 类客户群的指标 A 的值低,V 值在这四类中属于最高,这说明该类客户群在遇到感兴趣的商品时不习惯于加购,而更习惯于将商品收藏。根据以上分析,可将这类客户归为中等价值客户群。这类客户有较大的价值挖掘潜力,公司应尽可能保持这些客户,提高这类客户在该平台的消费频率,使他们不能够容易地转向竞争对手。

第三类客户群为 D 类客户,包括 16 575 名客户,占该电子商务平台数据集的 44%。与其他三种客户类型的指标值相比:这类客户的购买频率

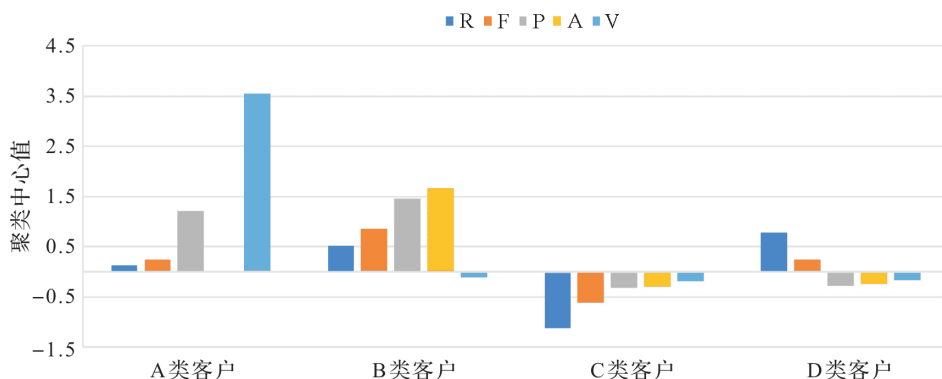


图4 RFAV模型各指标分布图

Figure 4 Distribution chart of RFAV model

属于中等程度,但是R值最高,说明很久没有在该电商平台购买商品;P值、A值和V值都较低,表明这类客户属于该电商平台中不活跃的一类群体,不会经常在该平台浏览商品、加购或者收藏。可以将其归类为低价值客户群。但是虽然D类客户群的当前价值一般但增值潜力较好,商家应该重点培养这类客户,增加个性化推送商品的频率,尽可能提升这类客户群的价值。

第四类客户群为C类客户,包括14156名客户,占该电子商务平台数据集的38%。与其他三种客户类型的指标值相比:这类客户的R值很低,但是F值也是最低的,说明这类客户群最近有在该平台消费,但总体消费频率很低;在P、A、V值方面类似于第三类客户,也属于不活跃的客户群。与D类客户不同的是,虽然这类客户的当前价值很低,购买频率和活跃度都一般,但他们最后一次消费日子却很近,所以这类客户很可能是新获得的客户。对这类客户需特别注意,应及时地推送客户可能感兴趣的物品,做好相关的服务,促进客户关系进一步发展,进而增加客户的购买频率,减少客户流失。

4 结 论

客户关系管理是电商平台保持竞争优势的前提。企业要实现高效的客户关系管理,必须提高客户细分的准确性和效率。针对现有研究存在的客户细分指标单一的问题,以及传统K-means人为设定K值的缺陷,在指标选取上,本文加入了客户的消费行为数据,选取RFAV作为细分指

标;在方法上,引入CH指标对传统K-means算法进行改进,并利用某电子商务平台的37376个客户样本进行实证研究。

基于本文的对比实验结果和实证结果,可以得到以下两点结论:第一,针对电子商务客户特点提出的RFAV客户细分指标是有效的,本文选取的客户价值特征和客户消费行为特征,能够识别和区分电子商务客户在进行线上消费时表现出的不同消费习惯和偏好;第二,改进K-means算法能够提高电子商务客户数据集的聚类准确性和效率,由于传统的K-means算法存在人为设定K值的缺陷,本文加入了CH聚类质量评价指标,对K-means算法进行改进。

当然,本文的研究也存在一定的局限性:第一,客户细分指标需要更加多元,由于电商平台的数据量庞大,数据维度多样,未来可以挖掘更加多元化的特征数据;第二,由于每种算法都有其自身的局限性,为了更好契合电商平台客户样本数据集,实现更有效的客户细分,未来可以考虑将密度聚类、层次聚类等方法引入电子商务客户细分研究中。

【参 考 文 献】

- [1] CHIANG W Y. To mine association rules of customer values via a data mining procedure with improved model: An empirical case study[J]. *Expert Systems with Applications*, 2011, 38(3): 1716-1722.
- [2] SOLTANI Z, NAVIMPOUR N J. Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future re-

- search[J]. *Computers in Human Behavior*, 2016, 61(3): 667-688.
- [3] 贾应丽. “大数据”背景下的客户关系管理研究——以 B2C 电子商务企业为例[J]. *电子商务*, 2018, 218(2): 7-10, 12.
- JIA Y L. Research on customer relationship management in the context of “big data”—a case study of b2c e-commerce enterprises[J]. *E-Business Journal*, 2018, 218(2): 7-10, 12.
- [4] 贺仁龙, 叶宇航, 赵晓康. 电信企业客户细分模型构建与应用研究[J]. *情报杂志*, 2011, 30(S1): 148-151.
- HE R L, YE Y H, ZHAO X K. Research on the construction and application of customer segmentation model in Telecom enterprises [J]. *Journal of Intelligence*, 2011, 30(S1): 148-151.
- [5] 李为康, 杨小兵. 一种改进的 RFM 模型在网店客户细分中的应用[J]. *中国计量大学学报*, 2020, 31(1): 85-91, 134.
- LI W K, YANG X B. Application of an improved RFM model in customer segmentation of online stores[J]. *Journal of China University of Metrology*, 2020, 31(1): 85-91, 134.
- [6] ROMDHANE L B, FADHEL N, AYEB B. An efficient approach for building customer profiles from business data [J]. *Expert Systems with Applications*, 2010, 37(2): 1573-1585.
- [7] CHOU P B, GROSSMAN E, GUNOPULOS D, et al. Identifying prospective customers [C]//*Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris: ACM, 2000: 447-456.
- [8] 刘英姿, 吴昊. 客户细分方法研究综述[J]. *管理工程学报*, 2006, 20(1): 53-56.
- LIU Y Z, WU H. A summarization of customer segmentation methods[J]. *Journal of Industrial Engineering and Engineering Management*, 2006, 20(1): 53-56.
- [9] WANG L. The impact of perception difference on channel conflict: A customer relationship life cycle view[J]. *Journal of Service Science and Management*, 2015, 8(5): 655-661.
- [10] MARTIN C, ADRIAN P, DAVID B. Relationship marketing[J]. *Butter Worth-Heinemann Ltd*, 2008, 32(4): 23-29.
- [11] BUCKINX W, VERSTRAETEN G, VAN D P D. Predicting customer loyalty using the internal transactional database[J]. *Expert Systems with Applications*, 2007, 32(1): 125-134.
- [12] CHIANG W Y. Discovering customer value for marketing systems: An empirical case study[J]. *International Journal of Production Research*, 2017, 55(17): 5157-5167.
- [13] PEKER S, KOCYIGIT A, EREN P E. LRFMP model for customer segmentation in the grocery retail industry: A case study[J]. *Marketing Intelligence & Planning*, 2017, 35(4): 544-559.
- [14] SARVARI P A, USTUNDAG A, TAKCI H. Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis[J]. *Kybernetes*, 2016, 45(7): 1129-1157.
- [15] SONG M, ZHAO X, HAI H E, et al. Statistics-based CRM approach via time series segmenting RFM on large scale data[J]. *Knowledge-Based Systems*, 2017, 132(15): 282-291.
- [16] 任春华, 孙林夫, 吴奇石. 基于 LRFAT 模型和改进 K-means 的汽车忠诚客户细分方法[J]. *计算机集成制造系统*, 2019, 25(12): 3267-3278.
- REN C H, SUN L F, WU Q S. Automobile loyalty customer segmentation method based on LRFAT model and improved K-means clustering[J]. *Computer Integrated Manufacturing Systems*, 2019, 25(12): 3267-3278.
- [17] HOSSEINI S M S, MALEKI A, GUOLAMIAN M R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty[J]. *Expert Systems with Applications*, 2010, 37(7): 5259-5264.
- [18] CHIANG W Y. Establishing high value markets for data-driven customer relationship management systems[J]. *Kybernetes*, 2019, 48(3): 650-662.
- [19] LI H, YANG X S, XIA Y, et al. K-LRFMD: Method of customer value segmentation in shared transportation filed based on improved K-means algorithm[J]. *Journal of Physics Conference Series*, 2018(1): 1-9.
- [20] HUGHES A. *Strategic Database Marketing: The Master Plan for Starling and Managing A Profitable Customer-Based Marketing Program* [M]. New York: McGraw-Hill Professional, 1994: 231-238.
- [21] 许学国, 桂美增. 基于深度学习的技术预测方法——以机器人技术为例[J]. *情报杂志*, 2020, 39(8): 53-62.
- XU X G, GUI M Z. Technology forecast based on deep learning—using robot technology as an example[J]. *Journal of Intelligence*, 2020, 39(8): 53-62.