云南民族大学学报(自然科学版),2024,33(2):251 - 257　　CN 53 - 1192/N　ISSN 1672 - 8513
doi:10.3969/j.issn.1672 - 8513.2024.02.014　　http://ynmz.cbpt.cnki.net

基于特征分箱和 *K* - Means 算法的用户行为分析方法

殷丽凤,路建政
(大连交通大学 软件学院,辽宁 大连 116028)

摘要:针对网购用户所产生的购物行为进行分析,首先通过数据处理构建客户关系管理模型(RFM 模型),在此模型的基础上采用特征分箱法和 *K* - Means 聚类两种方法对用户进行细分,并对 2 种模型结果进行比较分析,讨论二者的差异性和具体的应用范围和意义. 其中,基于特征分箱法的 RFM 模型将变量转化到相似的尺度上并将变量离散化,使得用户分类标签更加清晰,也可依据各类标签分类出不同类型的用户. *K* - Means 算法通过轮廓系数评估聚类算法质量以至于选取最优 K 值. 本文实验分析结果可为运营商提供更加可靠直观的数据,使得运营商可以根据不同用户的不同行为进行市场细分,进而进行精准营销和服务设置.

关键词:特征分箱;*K* - Means 算法;用户行为;RFM 模型;网购
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1672 - 8513(2024)02 - 0251 - 07

随着互联网行业的蓬勃发展,网上购物已经深入千家万户,人们足不出户就可以买到自己喜欢的物品. 但是,用户在购物过程中所产生的行为是千变万化的. 对于网络购物运营商而言,如何将错综复杂的用户行为数据通过数据处理的方式进行数据分析分析以此来评估和维护客户关系,满足客户日益增长的个性化需求是整个电商平台决策运营系统中的基础.

客户关系管理(CRM)成为最主流的"以客户为中心"的管理理念之一以满足业务发展的需要. 客户分类作为 CRM 的重要管理工具,是它的重要依据,是营销的重要依据[1-2]. 徐翔斌等[3]通过引入客户关系管理(recency frequency monetary,简称 RFM)模型总利润属性对电子商务客户进行分类. HU Y 等[4]提出将高进度、高频次、高消费价值的客户定义为高价值客户,然后根据 RFM 模型估算客户价值,以有效提高现有客户的价值转化率,但他并没有提出相应的比较算法. 陈子璐[5]建立 RFM 模型,利用 *K* - Means 算法和四分位法对客户进行细分,比较两种方法的优缺点,帮助平台根据实际需求选择合适的方法. 寻找最佳和潜在客户的方法,实施有针对性的策略吸引客户,形成长期购买行为,提高客户忠诚度. 她虽然比较了 2 种方式,但并没有考虑到公司的实际情况. 程汝娇等[6]提出了 1 种基于 RFM 模型的半监督聚类算法,该算法在传统 *K* - Means 算法的基础上采用自适应方法确定 K 值和初始聚类中心. 引入必须链接和不能链接约束将类别标签转换为成对约束信息. 提出基于 HMRF - *K*Means 成对约束,引入约束惩罚和约束奖励条件来调整聚类指导和聚类结果. 陈东清等[7]提出基于熵方法改进 RFM 模型的电子商务客户价值分割研究. 李伟康等[8]引入了 RFM 模型,以平均交易间隔、平均消费金额和平均产品浏览量作为重要属性,并应用层次分析优化顾客细分. 李斌等[9-10]将聚类分析数据挖掘技术应用于客户关系管理,可以改善客户关系,预测未来趋势和行为,为决策提供支持. 使用最小方差方法的谱系聚类对样本数据进行聚类,挖掘分析客户群体中具有不同特征的群体,得到直观的聚类过程和更合理的分组结果. 吴涛[11]使用 RFM、*K* - Means 和 *K* - Means + + 分析客户上次购买日期与当前日期的间隔;客户在一定时间内的购买次数;客户在一定时间内的消费总量. 利用 3 种客户行为指标对 3 种方法进行分类评价. 许雪晶等[12]采用 RFM 模型结合 *K* - Means 聚类算法对公司 2018 年 28 162

收稿日期:2022 - 03 - 09.
基金项目:国家自然科学基金(61771087).
作者简介:殷丽凤(1976 -),女,博士,副教授. 主要从事大数据挖掘、大数据分析、不确定 XML 规范化处理、查询研究.

笔订单交易数据进行聚类,并对其进行评估、分析和研究. 蒋伟等[13]使用2层RFM模型研究了黔彩新新零售会员价值. 这种模式不仅可以按价格对会员进行细分,还可以根据营销需要,在消费频次和消费金额上进一步细分. 结合会员生命周期管理,可为管理者制定营销策略、提升会员价值提供可靠、具体、科学的指导.

以上学者大多只做实验分析,没有考虑什么样的方法更适合企业自身的实际发展. 本文将在RFM模型的基础上采用数据挖掘技术中的特征分箱法构建新的RFM模型和经典聚类算法 K – Means 2种方法对用户进行细化分组,并对两种模型进行比较,讨论二者的优缺点和具体的应用范围及意义. 并根据实验分析结果评估用户的类型和价值. 针对不同用户群体采取针对性的营销策略,进而提高用户对电商平台的满意度,使得商家挖掘更多有价值的客户,提高市场竞争力.

1 基本方法

1.1 K – Means 算法

K – Means 算法[14]是聚类算法中的一种经典算法,同时也被称为是一种基于形心的算法技术. 它的处理过程如下:首先,在数据集D中随机选取k个对象,每个对象代表一个簇的初始类中心. 对剩下的每个对象,计算其与各个簇中心的欧式距离,并将此对象归类到最相近的簇中. 然后,K – Means 算法不断迭代改善簇内对象差值. 针对每个簇而言,它使用上述迭代归类到该簇的对象,再计算新的均值或中心. 然后,使用更新后的均值作为新的簇中心,再将所有对象重新归类. 迭代继续,直到分配稳定,最终更新的簇与前一轮更新的簇相同.

本文采用欧式距离,簇中所有样本点到聚类中心之间的误差的平方和,定义为:

$$E=\sum_{i=1}^{k}\sum_{p\in C_i} dist(p,c_i)^2. \tag{1}$$

其中,E 是数据集中所有对象的误差和平方和;P 是空间中的点,表示给定的数据对象;c_i 是簇 C_i 的形心(P 和 c_i 都是多维的). 换言之,对于簇中的每个对象,求对象到其簇中心距离的平方,然后求和[15-16]. 这个目标函数试图使生成的结果簇尽可能紧凑和独立.

1.2 RFM 模型

客户关系管理模型(RFM模型)是根据客户活跃程度和交易金额的贡献,进行客户价值细分的一种方法.

RFM模型[17]最初是由Hughes于1994年提出,它包括 R(recency)、F(frequency)、M(monetary)3个变量. R 表示最近一次交易时间间隔. 基于最近一次交易日期计算的得分,距离当前日期越近,得分越高;F 表示客户最近一段时间内的交易次数. 基于交易频率计算的得分,交易频率越高,得分越高;M 表示客户最近一段时间内的交易金额. 基于交易金额计算得分,交易金额越高,得分越高. RFM总分值公式如下:

$$\mathrm{RFM}=\omega_R\times R+\omega_F\times F+\omega_M\times M. \tag{2}$$

其中RFM指客户的综合RFM值,ω_R、ω_F、ω_M 分别表示 R、F 和 M 在计算客户价值的权重[18].

1.3 轮廓系数

轮廓系数(silhouette coefficient)[19]是对聚类效果优劣的一种评价方式. 最早是由Peter J. Rousseeuw在1986年提出的,它将凝聚度和分离度两种因素相结合,使它可以在相同的原始数据的基础上评价不同的算法、或者不同的运行方式的算法对聚类结果所产生的影响.

根据样本 i 的簇内不相似度 a_i 和簇间不相似度 b_i,定义样本 i 的轮廓系数如下所示:

$$s_i=\frac{b_i-a_i}{\max\{a_i,b_i\}}. \tag{3}$$

其中 s_i 用来评价样本 i 是否合适于所在的簇,因为 s_i 的取值范围在 –1 到 1 之间,若 s_i 的值接近1,则表明簇内平均距离 a_i 小于最小簇间平均距离 b_i,则说明样本 i 聚类合理;反之,若 s_i 的值接近 –1,则说明样本 i 的聚类效果不太理想,样本 i 更适合聚类到其他簇中;如果 s_i 近似为0,则说明样本 i 在两个簇的边界上[20].

所有样本的 s_i 的平均结果称为聚类结果的轮廓系数,它是评价该聚类结果是否合理的有效度量.

2 用户购物行为分析

2.1 用户购物行为分析思路

电子商务数据中隐藏着巨大的商业价值,其中用户的购物行为数据隐藏了消费者的购物习惯和特征. 本文将重点通过数据挖掘来挖掘消费者的特征,并针对每个消费者的特征采用两种方法进行细化分组,为企业做精准营销提供参考.

本文首先在数据预处理后构建RFM模型. 构建RFM模型的目的是根据用户的活跃度和交易金额的贡献来提取用户购物行为的基本特征. 其次,在构建RFM模型后,采用等距分箱的方法对每个特征数据进行分组和离散,得到基于特征分箱的新的RFM模型. 这种

类型的模型使得用户的 RFM 模型特征分组更加详细和清晰. 然后在 RFM 模型的基础上进行 K – Means 聚类,使用聚类方法自动对用户进行分组. 最后对两种分组方式进行比较,从企业的实际情况出发,探讨两种算法的优缺点,为企业选择营销策略提供参考.

2.2　用户购物行为分析的模型与步骤

Step 1:进行数据预处理,提取与本次实验相关的实验数据并填写实验缺失的数据,然后处理一些影响实验结果的退货单和消费金额.

Step 2:构建 RFM 模型,计算三类用户特征数据,即用户最后一次购买的时间间隔(R)、消费频次(F)、消费总金额(M).

Step 3:构建基于特征分箱的 RFM 模型,将用户上次购买时间间隔、消费频率、消费总金额三类用户特征数据划分为等距分箱,对数据进行离散化和打分,最终得到新的 RFM 模型.

Step 4:构建 K – Means 模型,利用基于 RFM 模型的轮廓系数评估进行 K – Means 算法聚类,得到 K – Means 分组模型

3　实验结果与分析

3.1　数据资源和实验环境

实验数据来自 The UCI Machine Learning Repository. 该 Online Retail II 数据集包含在 2010/12/12 和 2011/12/09 之间发生在英国的注册非商店在线零售的所有交易. 该公司主要销售独特的礼品. 公司的许多客户都是批发商. 数据集包括 1067371 行数据和 8 个属性. 属性包括:Invoice(发票编号)、StockCode(产品编号)、Description(产品描述)、Quantity(产品数量)、InvoiceDate(发票时间)、Unit – Price(产品单价)、CustomerID(客户编号)、Country(国家).

本实验在单机 Windows11 系统上用 Python 语言编译,使用的 IDE 为 Pycharm.

3.2　参数设置

RFM 模型中“Recent”、“Frequency”和“Monetary”字段中的进行等距分箱,每个字段中的样本数据分为 5 个区间,分别进行评分. 最高分 5 分,最低分 1 分. 用户价值表按上述方法划分后,对数据进行离散化,选取“R_分数”、“F_分数”、“M_分数”三个字段中数据的平均值作为分割点,划分为“高”和“低”类别,分别记为“1”、“0”,以此得到基于特征分箱方法的 RFM 模型,最后建立细分规则如下:“111”:重要价值用户;“101”:重要发展用户;“011”:重要保留用户;“110”:普通值用户;001”:重要扣留用户;“100”:一般发展用户;“010”:一般用户;“000”:普通挽留用户. 在使用 K – Means 算法的实验中,K = 4 是最优的类别数.

3.3　数据预处理

首先检查数据集中各个字段是否存在缺失值,将与本次实验相关的字段中存在的缺失值进行填补,并删除与本次实验无关的字段. 其次,对退货订单的处理,在数据集 Invioce 标签中,存在一些包含字母“C”的订单编号,此类订单编号包含的含义是订单退货. 所以在数据预处理阶段要记录每一笔订单的退货量以便之后计算单笔订单的真实消费金额. 最后提取出和取消订单的商品相配对的原订单并删除 Quantity 字段中为负的订单. 其中各个月份退货订单的变化如图 1 所示:

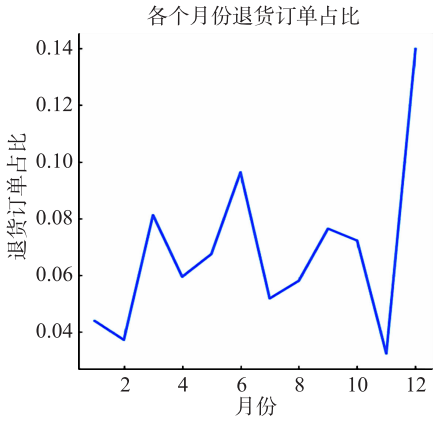

图 1　各月退货单占比

图 1 显示,1 月和 12 月的退货比例有所增加. 这里的原因应该是圣诞节的到来,导致大量购买后大量退货.

截取数据预处理之后的 5 类订单数据如表 1 所示:

表 1　数据预处理之后的部分数据

用户 ID	订单 ID	金额	交易时间
12346	491725	45.0	2009 – 12 – 14
12346	491742	22.5	2009 – 12 – 14
12346	491744	22.5	2009 – 12 – 14
12346	492718	22.5	2009 – 12 – 18
12346	492722	1.0	2009 – 12 – 18

3.4　基于特征分箱法的 RFM 模型的建立和分析

在预处理之后的数据中,订单金额是每位用户的单笔消费金额,所以需要将每位客户的单笔金额求和得到单位用户的总消费额;用户中最后订单时间与单个用户的最后订单时间的差值就是此用户最近一次购买的时间;单个用户的总的消费次数就是

此用户的消费频率. 经过以上分析和处理得到 RFM 用户价值的部分数据表如表 2 所示:

表 2 RFM 部分用户价值表

用户 ID	最近购买间隔(R)	消费频率(F)	消费额(M)
12346	164	11	368.36
12347	2	2	1 323.32
12348	73	1	222.16
12349	42	3	2 671.14
12351	10	1	300.93

将 RFM 模型中的"最近购买间隔(R)","消费频率(F)"和"消费金额(M)"三个字段中的数据进行等距分箱,其中各个字段中的样本数据分为 5 个区间,并分别为其打分. 最高分为 5 分,最低分为 1 分. 具体划分方法如表 3 所示:

表 3 特征分箱划分方法表

R_分箱	F_分箱	M_分箱	分数
[0,30]	(16,32]	(2 400,4 800]	5
(30,60]	(8,16]	(1 200,2 400]	4
(60,90]	(4,8]	(600,1 200]	3
(90,180]	(2,4]	(300,600]	2
(180,360]	[1,2]	[0,300]	1

其中"R_分箱"代表最近购买间隔的分段范围,"F_分箱"代表消费频率的分段范围,"M_分箱"代表消费金额的分段范围. 通过上述方法划分用户价值表之后再将数据离散化,选取"R_分数","F_分数"和"M_分数"3 个字段中的数据的均值作为分割点分为"高"和"低"2 类,分别记为"1"、"0". 最终得到基于特征分箱法的 RFM 用户价值表如表 4 所示:

表 4 基于特征分箱法的部分 RFM 用户价值表

用户 ID	R_分数	F_分数	M_分数	R	F	M	RFM 值
12346	2.0	4.0	2.0	0	1	0	010
12347	5.0	1.0	4.0	1	0	1	101
12348	3.0	1.0	1.0	1	0	0	100
12349	4.0	2.0	5.0	1	0	1	101
12351	5.0	1.0	2.0	1	0	0	100

根据特征分箱法得到的 RFM 用户价值表可以将客户进行细化分组,在特征离散化的基础上将用户细分为 8 个不同类型的客户,细分规则如下:

"111":重要价值客户;"101":重要发展客户

"011":重要保持客户;"110":普通价值客户

"001":重要挽留客户;"100":普通发展客户

"010":普通客户 ;"000":一般挽留客户

根据细分规则得到客户分组表如表 5 所示:

表 5 基于特征分箱的 RFM 客户分组表

用户 ID	R_分数	F_分数	M_分数	R	F	M	RFM 值	用户等级
12346	2.0	4.0	2.0	0	1	0	010	普通客户
12347	5.0	1.0	4.0	1	0	1	101	重要发展客户
12348	3.0	1.0	1.0	1	0	0	100	普通发展客户
12349	4.0	2.0	5.0	1	0	1	101	重要发展客户
12351	5.0	1.0	2.0	1	0	0	100	普通发展客户

经过分析和处理,得到了最终基于特征分箱的 RFM 客户价值分组表,根据客户价值表确定各类用户等级,并进行用户分层,得到各类用户占比. RFM 用户分层图如图 2 所示.

将基于特征分箱的 RFM 用户价值表(表 5)和 RFM 用户价值占比图(图 2)相结合得出结论:

满足"111"标签的用户占比 24.6%. 这类用户总的来说最近一次购买时间间隔较短,购买频次较高且消费金额较大,所以将此类用户划分为重要价值客户.

满足"100"标签的用户占比 24.3%. 这类用户最近一次购买时间间隔较短,但是消费频率和消费总金额较低,所以将此类用户划分为一般发展用户.

满足"000"标签的用户占 20.4%. 这类用户无论是消费频率与消费金额还是最近一次购买时间间隔得分都比较低,所以把此类用户划分为挽留客户.

满足"101"标签的用户占比 13.5%. 这类用户最近一次购买时间间隔较短并且消费的总金额较大,但是消费频率较低. 所以把此类用户划分为重要

发展客户.

满足“001”标签的用户占比 6.6%. 这类用户最近一次购买时间间隔较长且消费频率较低. 但是消费的总金额较大. 所以把此类用户划分为重要保持客户.

满足 “110”标签的用户占比 6.4%. 这类用户最近一次购买时间间隔较短且消费频率较低,但是总的消费金额较小. 所以把此类用户划分为普通价值客户.

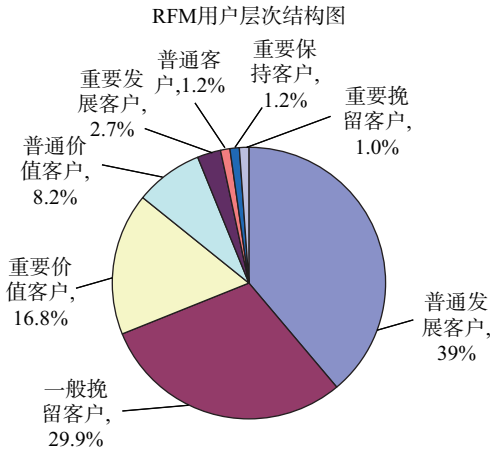

图 2　RFM 用户价值占比图

满足“011”标签的用户占比 3.2%. 这类用户消费频率和消费总金额都较高,但是最近一次购买时间间隔较短. 所以把此类用户划分为重要保持客户.

满足“010”标签的用户占比 0.9%. 这次用户最近一次购买时间间隔较低且消费总金额较低,但是消费频率较高. 所以把此类用户划分为普通客户.

3.5　基于 K－Means 算法的用户购物行为的分析

在 RFM 模型的基础上选取“R_score”、“F_score”、“M_score”作为 K－Means 算法的聚类变量并基于 Python 语言进行数据分析. 首先通过轮廓系数来评估算法的质量以确定最优 K 值. 其中轮廓系数曲线图如图 3 所示:

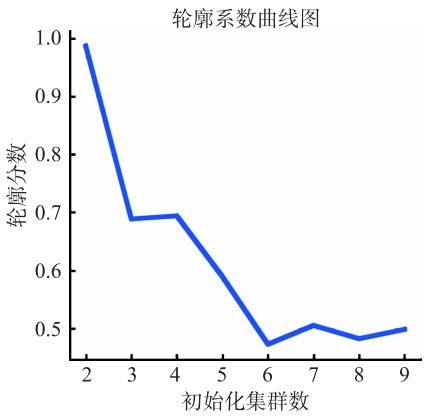

图 3　轮廓系数曲线图

根据轮廓系数评判规则判定轮廓系数得分越接近 1 样本聚类越合理,由上述轮廓系数曲线图(图 2)可知,当 K 取 2 时聚类效果最优,但是出于实际情况考虑,K = 2 时不符合实际要求,所以,当 K = 4 时,聚类效果较优. 进而得到聚类后的客户划分如表 6 所示:

表 6　RFM 模型用户价值表

分组	人数	R	F	M
0	1 208	－1.104 927	－0.850 492	－0.837 790
1	1 673	0.788 890	－0.154 506	－0.243 101
2	700	－0.911 078	0.179 349	0.340 983
3	722	0.904 008	1.607 119	1.634 446

根据聚类好的客户价值表得到可视化模型雷达图(Radar Chart),可以更直观的发现各个客户群的差异. 聚类可视化模型如图 4 所示:

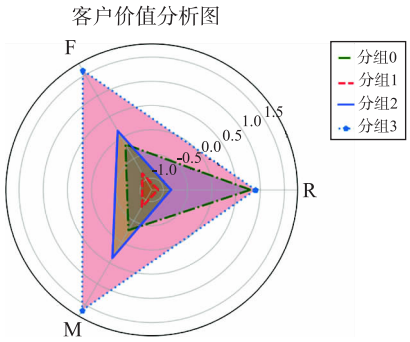

图 4　客户分组雷达图

结合表 6 和图 4 可以看出将最近购买时间间隔(R)较短,消费频率(F)较低且消费金额(M)较低的用户分为第一组(分组 0),此类用户约占统计总人数的 38.9%,根据此类用户的购买特征可划分为重要挽留客户.

将最近购买时间间隔(R)最长,消费频率(F)最低且消费金额(M)最少的用户分为第二组(分组 1),此类用户约占统计总人数的 28.1%,根据此类用户的购买特征可分为一般发展客户.

将最近购买时间间隔(R)较长,消费频率(F)较高且消费金额(M)较高的用户分为第三组(分组 2),此类用户约占统计总人数的 16.3%,根据此类用户的购买特征可划分为重要保持客户.

将最近购买时间间隔(R)最短,消费频率(F)和消费金额(M)最高的用户分为第四组(分组 3),此类用户约占统计总人数的 16.7%,根据此类用户的购买特征可划分为重要价值客户.

4 两种方法的对比分析

4.1 实验结果对比分析

2类方法虽然都可以对用户进行细化分组,但是二者之间仍存在差异性. 其差异性具体表现在两个方面:

细分程度上:基于特征分箱法的RFM模型将样本数据离散化后可以细分更多类别. 但是 K - Means 算法为考虑最优化会受到K值的限制使得类别数量也会存在一定局限性.

细分质量上:基于特征分箱法的RFM模型只是简单粗暴地把各个用户特征按照指定地阈值和区间进行细分,若两个用户的样本数据都接近两个相邻区间端点时,两位用户的性质相差无几. 但按照划分规则会把两位用户划分为不同类别的客户,进而一定程度上影响企业地营销判断. 然而,K - Means 算法聚类更加科学. 该算法通过不断优化迭代,在已经求得的聚类上再进行迭代修正确定部分样本的聚类,克服了少量样本聚类的不准确性,使得样本划分更合理.

无论选择哪种方法的用户群体划分都要考虑企业的实际能力,并且划分用户的不同类型目的也是帮助企业做到精准营销,所以,笔者认为两类方法的应用范围也有不同. 若公司资金储备充足,且营销体系完善则可以选择特征分箱的RFM划分方法,因为此类方法划分客户类别广泛,若能做到一对一营销管理,可以有效提高用户的满意度. 若公司资金相对短缺,且没有足够的人力物力投入营销,倘若多一个类别多一份负担的话,则可以选择 K - Means 聚类算法进行用户分组,虽然此类方法划分虽然收到K值限制,但是划分质量也有保障,可以为企业减少人力财力的投入,减轻起的营销成本.

4.2 营销策略分析

针对一般发展客户而言,这类客户平台可发送短信或者电子邮件进行找回召回,努力将这类用户转化为重要挽留客户或重要保持客户.

针对重要挽留客户而言,此平台也许并不是他们购物时的首选平台,或者对网购的需求不算很高. 针对此类用户,平台可实施问卷有礼的方式了解这类消费者的满意度并针对存在的不足适当改进,提升购物体验,增大用户粘性.

针对重要保持客户而言,平台可以推送一些相关其他商品,并做一些相应的促销和优惠活动,保持这类消费者的消费兴趣.

针对重要价值客户而言,电商平台需要大力挽留和着重维护. 并且还需要提高该部分用户的服务满意度,可适当发放一些福利. 并且值得注意的是,在做运营推广时要给予此类用户特别关注,避免引起此类用户的反感.

5 结语

在RFM模型的基础上,采用特征分箱方法构建新的RFM模型,并与 K - Means 算法进行比较,分析两种算法的优缺点和差异. 从企业实际出发,对两种方式的选择提出合理建议. 针对不同的用户群体进行详细分析,提出可行的营销策略.

参考文献:

[1] 马椿荣. 消费者价值研究理论综述[J]. 商业时代,2014(10):60-61.

[2] DUBOFF R S. Marketing to maximize profitability[J]. The Journal of Business Strategy,1992,13(6):10-13.

[3] 徐翔斌,王佳强,涂欢,等. 基于改进RFM模型的电子商务客户细分[J]. 计算机应用,2012,32(5):1439-1442.

[4] HU Y,YEH W. Discovering valuable frequent patterns based on RFM analysis without customer identification information[J]. Knowledge-Based Systems,2014,61(2):76-88.

[5] 陈子璐. 基于RFM模型的电子商务客户细分[J]. 市场周刊,2020(4):56-58.

[6] 程汝娇,徐鸿雁. 基于RFM模型的半监督聚类算法[J]. 计算机系统应用,2017,26(11):170-175.

[7] 陈东清,叶翀,黄章树. 基于熵权法改进RFM模型的电商客户价值细分研究[J]. 西安电子科技大学学报:社会科学版,2020,30(2):39-45.

[8] 李为康,杨小兵. 一种改进的RFM模型在网点客户细分中的应用[J]. 中国计量大学学报,2020,31(1):86-91.

[9] 李斌,郭剑毅. 聚类分析在客户关系管理中的研究与应用[J]. 计算机工程与设计,2005,26(2):540-542.

[10] 李斌,郭剑毅. 基于系统聚类的客户分析[J]. 昆明理工大学学报,2004,29(6):66-69.

[11] 吴涛. 基于RFM模型的电子商务顾客细分研究[J]. 铜陵学院学报,2020,19(5):55-59.

[12] 许雪晶,林辰玮. 基于RFM的电商数据客户价值细分实例研究[J]. 长春师范大学学报,2021,40(4):60-69.

[13] 蒋伟,嵩涛,罗恒. 基于双层RFM模型的黔彩新零售会员价值分析[J]. 内蒙古科技与经济,2020(19):62-63.

[14] 刘维. 数据挖掘中聚类算法综述[J]. 江苏商论,2018,7(30):120-123.

[15] HAN J W,KAMBER M. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 3 版. 北京:机械工业出版社,2012:293 –294.

[16] 黄晓辉,王成,熊李艳,等. 一种集成簇内和簇间距离的加权 K – Means 聚类方法[J]. 计算机学报,2019,42(12):2836 –2848.

[17] 陆娜,刘晓文,李兰. 基于 RFM 的网店客户价值细分研究[J]. 电脑知识与技术,2018,14(18):275 –276.

[18] 徐文瑞. 基于 RFM 模型的顾客消费行为为与顾客价值预测研究[J]. 商业经济研究,2017(19):44 –46.

[19] 朱连江,马炳先,赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用,2010(S2):139 –141.

[20] 尹世庄,王韬,谢方方,等. 基于互信息和轮廓系数的聚类结果评估方法[J]. 兵器装备工程学报,2020,41(8):207 –213.

User behavior analysis method based on feature binning and K – Means algorithm

YIN Li-feng ,LU Jian-zheng

(Software Technology of Dalian Jiaotong University,Dalian 116028,China)

Abstract:To analyze the shopping behavior generated by online shopping users,firstly,a customer relationship management model (RFM model) is constructed through data processing,and on the basis of this model,two methods of feature binning and K – Means clustering are used to classify users ,and compare and analyze the results of the two models,discuss their differences and specific application scope and significance. Among them,the RFM model based on the feature binning method converts the variables to similar scales and discretizes the variables,so that the user classification labels are clearer,and different types of users can also be classified according to various labels. The K – Means algorithm evaluates the quality of the clustering algorithm by the silhouette coefficient so as to select the optimal K value. The experimental analysis results in this paper can provide operators with more reliable and intuitive data,so that operators can segment the market according to the different behaviors of different users,and then conduct precise marketing and service settings.

Key words:feature binning;K – means;user behavior;RFMmodel;online shopping

(**责任编辑**　段　鹏)