文章编号：2095-2163(2023)09-0194-04 中图分类号：TP311.13；F274 文献标志码：A

优化的 K-means 聚类算法在客户细分中的应用研究

唐 欣
（北方民族大学 数学与信息科学学院，银川 750021）

摘 要：传统的 K-means 聚类算法虽然操作简单快捷，但因随机选取聚类中心等问题容易陷入局部最优，导致算法不稳定。本文从样本间的关系出发，利用样本密度来优化 K-means 算法，并利用聚类有效性指标进行比较，优化后的 K-means 算法更具有稳定性且聚类准确率更高。最后，将该算法应用到客户细分 RFM 模型中，依据聚类结果找到适合不同消费者的营销策略，从而帮助企业更好地为其提供差异化、个性化服务。

关键词：K-means 算法；样本密度；客户细分；RFM 模型

Research and application of customer segmentation model based on optimized K-means algorithm

TANG Xin
（School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China）

【Abstract】 Although the traditional K-means clustering algorithm is simple and fast, it is easy to fall into local optimization due to the random selection of clustering centers, which leads to instability of the algorithm. This paper proposed an optimized K-means algorithm with the relationship of datasets and compared it with traditional K-means by clustering effectiveness index. It is concluded that the optimized K-means algorithm is more stable and has higher clustering accuracy. Finally, this algorithm is employed to the customer segmentation RFM model and give different market strategies with different customers by the clustering results. Thus, it is much helpful for companies to provide differentiated and personalized services for them.

【Key words】 K-means clustering; sample density; customer segmentation; RFM model

0 引 言

互联网技术的快速成长，带动了电商、教育、医疗以及生物科技等领域的不断突破创新，大数据成为生活中不可或缺的一部分，使得交通出行、网上购物、线下支付等一系列活动简便快捷。信息多元化，数据挖掘与获取信息密不可分，通过数据清洗、转换以及集成等方式来挖掘有效信息。聚类分析是数据挖掘常用的聚类方法，利用同一类簇相似性高，不同类簇相似性低的行为准则划分数据，市场研究人员也常常将这一方法运用到客户细分中。

20 世纪 50 年代中期，美国学者温德尔史密斯根据市场细分准则提出了客户细分的概念[1]。即基于某一标准，将企业库中的所有客户划分为多种类型的客户群的过程[2]。利用聚类分析对客户划分的方法，能够挖掘更多有用信息，帮助企业了解客户的消费行为、习惯以及购物偏好等相关信息，达到更好地为客户提供个性化、差异化服务与体验的目的，进而有针对性地制定营销策略，促进公司持续健康发展。

不同的企业往往会制定不同的客户细分准则，挖掘客户特点，建立与客户之间的联系，实现公司利益最大化。Wang L 等[3]人从生命周期的角度出发，认为客户在生命的不同阶段会产生感知差异，而这种差异往往会带来不同的消费行为；王璀璨等[4]从客户价值的角度出发，优化客户关系管理系统，实现对电商企业的客户细分；Hughes 等[5]从客户行为的角度出发，通过建立 RFM 模型（R 代表最近一次消费时间、F 代表消费频率、M 代表消费总金额），了解客户消费行为习惯，做出不同价值分类，为企业更加有针对性地管理提供新思路。根据八二法则可知，企业的 80%利润往往是来自于 20%的忠诚客户[2]，说明了利用 RFM 模型，通过给出不同客户的价值分类，将这部分客户转化为忠诚客户后，他们重复购物

作者简介：唐 欣(1998-)，女，硕士研究生，主要研究方向：三支决策、聚类分析、数据挖掘等。

收稿日期：2022-09-13

能力往往能为企业带来更多的利润来源，而维持这部分客户远远小于获取一个新客户所要花费的成本。

很多学者为了更好地探索客户细分模型，常常利用数据挖掘的手段，结合聚类分析的方法来对客户进行划分。原慧琳等[6]从微观和宏观两个角度出发，利用 K-means 聚类算法对零售会员数据进行特征划分；杨琳等[7]根据民航客户自身特点，结合聚类分析方法，对 RFM 模型进行了改进，进一步提高了民航企业的服务质量；闫春[8]等利用轮廓系数改进 K-means 选取聚类数目，并在寿险数据中为挽留高价值客户提供了较高的决策依据。因此，将聚类分析方法应用到不同类型的客户群的划分中，能够帮助企业了解不同客户需求，给出客户价值定位，重新构建客户管理体系，提供个性化服务。本文将优化的 K-means 算法应用到 RFM 模型中，实现对客户数据的聚类，并根据聚类结果找出企业库中的忠诚客户，从而有效制定营销策略。

1　相关知识

1.1　K-means 算法

在聚类分析中，K-means 聚类是最常见的一种数据挖掘算法，是由 Macqueen 提出来的基于划分的聚类方法[9]。K-means 算法的聚类速度快，操作简单快捷，但聚类过程也存在一些缺陷，如依赖初始聚类中心的随机选取、极易受异常值影响、聚类结果不稳定等[10]。该算法通常使用欧氏距离来作为衡量两个对象之间的相似度指标，划分聚类结果，其基本思想是选择任意的 k 个初始聚类中心，计算出剩余数据对象与聚类中心的欧氏距离，找到距离最近的 $k-1$ 个聚类对象，不断更新迭代聚类中心，直到误差平方和(SSE)，即准则函数收敛，得到聚类结果，表示为 $C=\{C_1,C_2,\cdots,C_k\}$。

假设有 n 个 m 维属性的数据集 $U\in\{x_p\}(p=1,2,\cdots,n)$，记 $c_i(i=1,2,\cdots,k)$ 为 k 个聚类中心，每个聚类中心 c_i 都有 m 维属性，记为 $c_{ij}(j=1,2,\cdots,m)$，则每个对象 x_p 距离每个聚类中心 c_i 的欧式距离定义为式(1)：

$$Dist(x_p,c_i)=\sqrt{\sum_{j=1}^{m}(x_{pj}-c_{ij})^2} \tag{1}$$

误差平方和定义为式(2)：

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i}|Dist(x,c_i)|^2 \tag{2}$$

1.2　优化的 K-means 算法

K-means 作为无监督学习算法的一种，不需要提前知道聚类类别，能够对无标识的对象进行聚类。利用欧氏距离做相似度度量指标，得出相同一类簇的距离越小，其相似度越高；不同类簇的距离越大，其相似度越低。该方法在聚类过程中会因为受到极端值的影响而改变类簇的紧密性与离散性，降低整个聚类的准确性。本文从样本间的关系出发，首先采用高密度代替距离均值的方式，利用公式(3)计算最近邻密度选取数据集中样本密度较高的点作为第一个初始聚类中心 c_1，将其余对象划分到已确定聚类中心的类别当中。

$$\bar{X}^*=\frac{1}{n}\sum_{p=1}^{n}x_p \tag{3}$$

其中，$x_p(p=1,2,\cdots,n)$ 表示 n 个 m 维数据集。

其次，在剩下的没有被划分类别的对象中，采用欧氏距离，利用公式(4)和公式(5)找到离 c_1 最远的下一个临时聚类中心点 c_k 并聚类：

$$c_i=\max_{1<p<q<n}\{Dist(x_p,x_q)\} \tag{4}$$

$$Den_\lambda(x_{p_i},c_i)=\{x_{p_i}\mid Dist(x_{p_i},c_i)\leqslant\lambda\} \tag{5}$$

其中，$Den_\lambda(x_{p_i},c_i)$ $(p_i=1,2,\cdots,n_i)$ 表示与聚类中心 c_i 距离小于 λ 的所有数据对象；n_i 表示 $Den_\lambda(x_{p_i},c_i)$ 中的对象数目；A 为任意正常数。

λ 的计算公式为

$$\lambda=A*Dist(x_{p_i},\bar{X}^*) \tag{6}$$

利用公式(7)计算该临时聚类中心的密度，搜索距离其密度平均值最近的数据对象作为更新的聚类中心；如此迭代，直到得到所有初始聚类中心。得到所有数据对象的二支聚类结果，表示为 $C'=\{C_1',C_2',\cdots,C_k'\}$。

$$\overline{Den_\lambda(x_{p_i},c_i)}=\frac{1}{n_i}Den_\lambda(x_{p_i},c_i) \tag{7}$$

K-means 聚类算法通常是任意选取 k 个聚类中心，通常带有一定的随机性。本文提出了一种优化选取初始聚类中心的方法，利用样本分布信息，选择密度最高的点作为初始聚类中心，有效解决了人为因素干扰或者极端值影响导致聚类陷入局部最优的问题；限制聚类对象在 λ 的范围内，更加精准地远离了噪声点的干扰；最终聚类结果中能够满足同一类簇的相似程度最高，不同类簇的相似程度最低的条件，确保了聚类的稳定性。

2　优化的 K-means 算法的实验验证

实验环境：Intel，CPU16 GB 内存，512 GB 固态

硬盘,Windows10 操作系统,开发工具是 Python3.8。

2.1 数据集选取

本文从 UCI(University of CaliforniaIrvine)数据集中选取了 5 组真实数据集,实验数据集描述见表 1。

表 1 实验数据集描述

Tab. 1 Dataset description of experiment

数据集	样本个数	属性个数	类别数
Bupa	345	6	2
Glass	214	9	6
Wine	178	13	3
Breast	106	9	6
Bank	1 372	4	2

2.2 实验结果与分析

在实际聚类的过程中,为了确保数据的准确性,在聚类之前对数据均采取了无量纲化处理。同时,本文利用两个聚类有效性指标:内部指标 Davies-Bouldin-Index(DBI)及外部指标 Accuracy(ACC),验证本文提出的优化后的 K-means 算法的聚类有效性。DBI 指标是通过数据对象之间的紧密程度和分离程度来判断其内部结构和分布状态,DBI 越小,说明同一类的相似性越高,不同类的相异性越高;ACC 指标是比较最终聚类结果与数据集原始真实标签值,从而判断数据的准确性。实验结果见表 2,

表 2 UCI 数据集上的实验结果

Tab. 2 Experiment results of UCI dataset

数据集	算法	DBI	ACC
Bupa	K-means	1.800 9	0.585 5
	本文算法	**0.664 0**	**0.740 7**
Glass	K-means	0.962 5	0.598 1
	本文算法	**0.682 5**	**0.883 1**
Wine	K-means	1.305 3	0.955 0
	本文算法	**0.471 7**	**0.959 6**
Breast	K-means	0.882 6	0.773 5
	本文算法	**0.694 7**	**0.955 6**
Bank	K-means	1.191 3	0.575 8
	本文算法	**0.253 4**	**0.988 3**

由表 2 可知本文算法在数据集上拥有更小的 DBI,说明优化后的 K-means 算法同一类簇之间的紧密性高,不同类簇之间的分离性高;同时,本文算法在数据集上均拥有了较高的准确率,聚类效果较好。说明表明优化的 K-means 聚类算法更具有有效性,将其应用到客户细分模型中去,可实现更好的聚类结果。

3 优化算法在客户细分中的应用

3.1 模型构建

本文利用 Kaggle 竞赛平台中下载的 2011~2014 年全球消费数据样本“Global Superstore”,选择了5 191条美国 “Business-to-Customer”领域的消费数据,消费时间为 2011 年 1 月 4 日至 2014 年 12 月 31 日。

首先,将数据进行预处理。对给出的5 191条消费者数据进行订单编号、日期、金额等指标进行筛选,样本均在正常范围内,无异常数据。

其次,创建 RFM 模型,得到一份只含有 R、F、M 3 个指标 409 * 3 的消费者数据(其中,R 表示消费者最近一次交易时间距离 2014 年 12 月 31 日的天数,F 表示消费者在这 4 年内的消费总频次,M 表示消费者在这 4 年内的消费总金额),部分数据见表 3。

表 3 消费者的 RFM 指标数据(部分)

Tab. 3 The RFM indicator data of consumers(partly)

客户 ID	R 值	F 值	M 值
AA-10645	55	6	5 073.975
AH-10210	6	9	4 805.344
AB-10105	41	10	14 473.571
SE-20110	9	11	12 209.44
ZC-21910	54	13	8 025.707
RB-19360	96	6	15 117.34

对 409 位消费者数据的 RFM 模型进行描述性统计分析见表 4。

表 4 消费者的描述性统计

Tab. 4 Descriptive statistics of consumption data

	R 值	F 值	M 值
Count	409	409	409
Mean	143.38	6.33	2 839.58
Std.	178.97	2.55	2 586.05
Min	0	1	4.83
Max	1 098	17	15 117.34

通过表 4 可知,R、F、M 3 个指标之间存在较大的差异性,为了避免 3 个指标的量级不同而影响到聚类结果,本文采用“Z-core 标准化”的方式处理数据,降低不同指标之间的差异,确保指标之间的变量具有可比性,标准化后的消费者的 RFM 数据见表 5。