doi:10.16104/j.issn.1673-1891.2022.01.017

K-Means 算法与数据挖掘在旅游电商平台设计中的应用

尹寿芳,张善智

(安徽工业经济职业技术学院商贸学院,安徽 合肥 230051)

摘　要:为了提升旅游电商服务水平,基于 K-means 聚类算法构建旅游电子商务平台,并采用随机梯度下降算法、自适应梯度优化算法和密度法对 K-means 聚类算法进行优化改进。结果表明:改进 K-means 聚类算法的系统响应速度相较于传统 K-means 聚类算法提升了 31.2%,电商平台推荐流量转化率为 2.93%,浏览行为中的推荐浏览率为 28.21%,购买行为中的推荐购买率为 15.37%,优于 Apriori 算法和 User-based CF 算法。利用改进 K-means 聚类算法构建旅游电子商务平台,能为平台用户提供个性化的旅游产品推荐,有效提升旅游产品的购买成交量,对旅游电商平台竞争力提升具有一定的实用价值。

关键词:数据挖掘;K-means 聚类算法;旅游;电子商务

中图分类号:F724.6;F592.6　　**文献标志码**:A　　**文章编号**:1673-1891(2022)01-0092-05

Application of K-Means Algorithm and Data Mining in the Design of Tourism E-Commerce Platform

YIN Shoufang,ZHANG Shanzhi

(Business College,Anhui Vocational College of Industrial Economics,Hefei,Anhui 230051,China)

Abstract:In order to improve the service level of tourism e-commerce,a tourism e-commerce platform is constructed based on K-means clustering algorithm,and the K-means clustering algorithm is optimized by random gradient descent algorithm,adaptive gradient optimization algorithm and density method.The experimental results show that the system response speed of the improved K-means clustering algorithm is 31.2% higher than that of the traditional K-means clustering algorithm,the recommended traffic conversion rate of e-commerce platform is 2.93%,the recommended browsing rate in browsing behavior is 28.21%,and the recommended purchase rate in purchasing behavior is 15.37%,which are better than the results by Apriori algorithm and user based CF algorithm.Using K-means clustering algorithm to build tourism e-commerce platform can provide personalized tourism product recommendation for platform users,effectively improve the purchase and trading volume of tourism products,and has important practical value for improving the competitiveness of tourism e-commerce platform.

Keywords:data mining; K-means clustering algorithm; tourism; electronic commerce

0 引言

近年来,随着国民经济水平的飞速增长,人们的生活质量和生活水平得到了显著提升,旅游消费在人们日常生活消费中所占的比重越来越高。电子商务行业的兴起和数据挖掘技术的发展为旅游服务业提供了新的消费模式,旅游电子商务平台已经逐渐成了人们获取旅游资讯和进行旅游产品预定的重要手段[1]。但大多旅游电商平台的商品查找流程烦琐,且产品推荐界面千篇一律,难以满足用户的需求。

近年来数据挖掘与电子商务结合的研究众多,研究多采用聚类分析、分类算法和关联分析 3 类数据挖掘方法进行电商精准营销。张磊等[2]利用 lightGBM 机器学习模型进行数据分类,挖掘电商广告转化率的影响因素,以此为基础对电商搜索广告进行优化调整,有效实现了电商平台的个性化广告推荐;郭艳萍[3]采用模糊运算聚类算法对电商客户数据进行数据信息挖掘分析,对电商平台用户进行聚类划分,为实现电商平台针对化服务提供辅助决

收稿日期:2021-11-08
基金项目:2020 年安徽省高校人文社科研究项目(SK2020A0090)。
作者简介:尹寿芳(1982—),女,安徽和县人,讲师,硕士,研究方向:电子商务、项目管理。

策;阿荣等[4]采用 Apriori 关联规则算法对电商平台用户进行分类,并根据用户的商品兴趣参数估计结果,为电商平台用户提供精准化客户服务。在以上研究中,数据挖掘与电子商务的有机融合已经取得了一定的成果,但所采用的数据挖掘算法仍存在一定的局限性,数据挖掘结果受数据集干扰因素影响较大,对多指标群体的划分精度不够高,需要进一步加强电商平台数据信息挖掘,为提升电商平台客户服务水平提供参考。

为了进一步提升旅游电商平台精准化营销的服务水平,本研究对 K-means 聚类算法进行优化改进,以提升 K-means 聚类算法的分析性能,并将其应用在旅游电子商务服务中,期望通过数据分析与整合的手段为用户定制个性化的旅游产品推荐界面,提升旅游电商平台的服务质量。

1 基于改进 K-means 聚类算法的旅游电商平台设计

1.1 基于 K-means 聚类算法的旅游电商平台

数据挖掘是一种基于数据库进行数据自动搜索的信息分析手段,通过对现有数据进行归纳整理和推理分析,挖掘数据中隐含的有价值的知识信息,分析整体趋势走向,从而对未来变化情况进行合理预测与决策[5-6]。以数据挖掘为基础构建旅游电子商务平台,通过对旅游电商平台用户相关信息数据进行智能化分析,对用户的消费潜力和消费倾向进行预测与判断,从而为用户提供具有针对化服务功能的旅游电商平台,在方便用户快速找到心仪的旅游产品的同时,提升旅游电商平台的销售量,增强旅游服务商的行业竞争力。利用旅游电商平台上用户的访问日志,挖掘用户的浏览偏好特征,根据不同的商品浏览属性特征,对旅游电商用户进行用户聚类,从而为用户提供个性化旅游商品推荐,优化用户的平台浏览体验,便于用户更快地寻找到满足自己需求的旅游产品。

聚类分析是常用的数据挖掘技术手段,根据数据对象之间的属性等联系,将数据库分为不同的类或簇,归属于同一类或簇的数据对象具有一定的相似性,通过相似度函数划分数据对象的相似性[7-8]。K-means 聚类算法通过聚类中心对数据对象进行聚类划分,随机选择 k 个聚类中心,按照就近原则将数据样本划分为 k 类,然后通过均值计算对归于同一类的数据样本进行聚类中心重新划分,反复进行聚类中心筛选操作,当聚类中心不再发生改变时算法终止,实现对数据对象的划分聚类,K-means 聚类算法运行流程如图 1 所示。

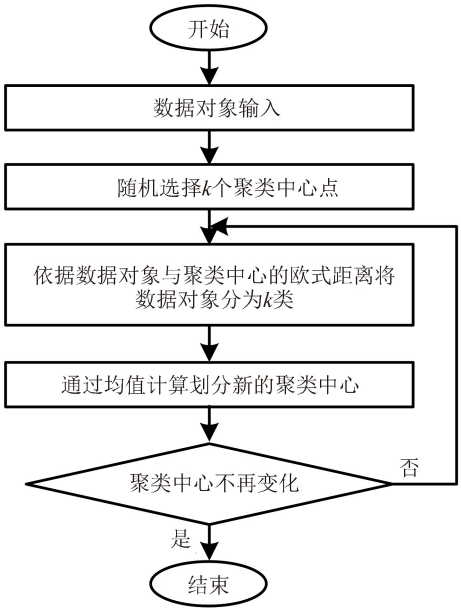

图 1　K-means 聚类算法运行流程

数据集合 $\boldsymbol{S}=\{x_1,x_2,\cdots,x_n\}$ 中包括 n 个 p 维的数据样本,数据集合的数据矩阵表示如下:

$$\boldsymbol{S}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

首先确定数据样本集合的聚类中心,随机选择 k 个聚类中心点,数据样本 x_i 与聚类中心的欧式距离计算函数 d 表示如下:

$$d(x_i,x_j)=(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2)^{\frac{1}{2}} \tag{2}$$

数据对象的差异性矩阵 $\boldsymbol{C}$ 表示如下:

$$\boldsymbol{C}=\begin{bmatrix} 0 & & & & \\ d(x_2,x_1) & 0 & & & \\ d(x_3,x_1) & d(x_3,x_2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(x_n,x_1) & d(x_n,x_2) & d(x_{2n},x_2) & \cdots & 0 \end{bmatrix} \tag{3}$$

通过均值计算的方式对不同类属的聚类中心点进行重新划分,得到聚类中心集合,算法反复迭代直到 k 个聚类中心不再发生变化。

1.2 K-means 聚类算法损失函数优化

传统 K-means 聚类算法的梯度更新以全量数据为更新对象,在面对海量数据聚类分析时,算法收敛速度较慢,更新计算量巨大[9-10]。为了提升 K-means 聚类算法的运行效率,采用随机梯度下降法进行梯度更新,利用样本的梯度值进行聚类中心的

更新操作,通过小部分样本的计算达到算法更新的目的,加快算法的收敛速度。将样本与最近聚类中心点的距离平方作为算法的损失函数,损失函数表示如下:

$$f(w)=\frac{1}{2\times(x-w^*)^2} \tag{4}$$

式中:w 表示损失值;x 表示数据样本集合中的一个随机样本;w^* 表示与该样本距离最近的聚类中心点。随机梯度下降法函数表示如下:

$$w=w-\mathrm{lr}\times(w^*-x) \tag{5}$$

式中:lr 表示学习率。利用随机梯度下降法对 w 进行更新,当聚类中心的变化值小于阈值或损失值变化小于阈值时,算法完成收敛,停止参数更新。

传统的 K-means 聚类算法较为复杂,容易发生过拟合现象,出现在训练数据集上表现较好而在测试集上表现较差的问题[11]。通常通过添加正则化项的方式防止过拟合现象的发生,对损失函数进行扩展,在损失函数中加入模型参数向量的范数,对模型复杂度进行惩罚[12]。将 L2 正则引入 K-means 聚类算法(L2 表示损失函数中模型参数向量的范数),求参数向量各元素的平方和,然后进行开方,利用 L2 范数提升算法损失函数的求解稳定性,避免算法过拟合。并利用 L2 范数将损失函数变为强凸函数,加快算法收敛,提升迭代的收敛效率。多项式模型出现过拟合现象时,其函数曲线与噪声点接近,出现在噪声点之间来回跳跃的情况,函数曲线部分区域的切线斜率较高,导致函数导数的绝对值过大[13]。L2 范数的引入可以使得较大参数的值均匀集中在 0 附近,有效提升算法的泛化能力,避免 K-means 聚类算法出现过拟合现象。

1.3 K-means 聚类算法学习率优化

学习率的选择影响算法的收敛速度,合适的学习率能有效提升算法的收敛效率,训练初期应采用较大的学习率来缩短训练时间,提升算法效率,训练后期应对学习率进行适当减小调整,避免出现因参数收敛速度较快而跳过极小值点的问题[14]。传统的 K-means 聚类算法利用固定的学习率值进行训练,容易造成算法震荡,影响算法性能,因此采用自适应梯度优化算法对学习率方向进行自适应确定,通过对历史梯度的指数衰减平均值和历史梯度平方的指数衰减平均值进行存储,自适应确定算法的衰减学习率。自适应梯度优化算法具有较强的鲁棒性,利用梯度的均值和有偏方差进行估计移动平均,通过偏差修正的方式减小初始化偏差,增强算法的实用性。一阶梯度和二阶梯度的衰减平均值计算函数表示如下:

$$\begin{cases} m_t=\beta_1 m_{t-1}+(1-\beta_1)\boldsymbol{g}_t \\ v_t=\beta_2 m_{t-1}+(1-\beta_2)\boldsymbol{g}_t^2 \end{cases} \tag{6}$$

式中:m 和 v 分别表示一阶动量和二阶动量;m_t 表示均值估计;v_t 表示有偏方差估计;β_1 表示一阶动量衰减系数;β_2 表示二阶动量衰减系数;t 表示初始化时间步长;$\boldsymbol{g}_t$ 表示偏导数向量。

在初始化的初期阶段,m 和 v 初始化为 0 向量,使得 m_t 和 v_t 会偏差向 0,对算法性能产生影响,因此在自适应梯度优化算法中增加偏差校正机制,通过偏差修正保证每次迭代学习率均保持在确定的范围内。均值估计和有偏方差估计修正函数表示如下:

$$\begin{cases} \hat{m}_t\leftarrow m_t/(1-\beta_1^t) \\ \hat{v}_t\leftarrow v_t/(1-\beta_2^t) \end{cases} \tag{7}$$

式中:$\hat{m}_t$ 表示校正后的均值估计;$\hat{v}_t$ 表示校正后的有偏方差估计;β_1^t 表示校正后的一阶动量衰减系数;β_2^t 表示校正后的二阶动量衰减系数。

采用自适应梯度优化算法对 K-means 聚类算法进行优化改进,通过指数衰减的方式进行学习率更新,从而控制梯度更新的步长,提升 K-means 聚类算法的收敛速度。

1.4 K-means 聚类算法初始中心点优化

传统 K-means 聚类算法对初始聚类中心点位置的依赖性较高,聚类中心点的初始位置直接影响算法的最终解的优劣,而传统 K-means 聚类算法的初始聚类中心点通过随机选择的方式确定,具有很强的不确定性[15]。因此采用密度法对 k 个聚类中心初始位置的选择方式进行优化,结合密度参数确定 k 个初始聚类中心,基于密度法的 K-means 聚类算法优化流程如图 2 所示。

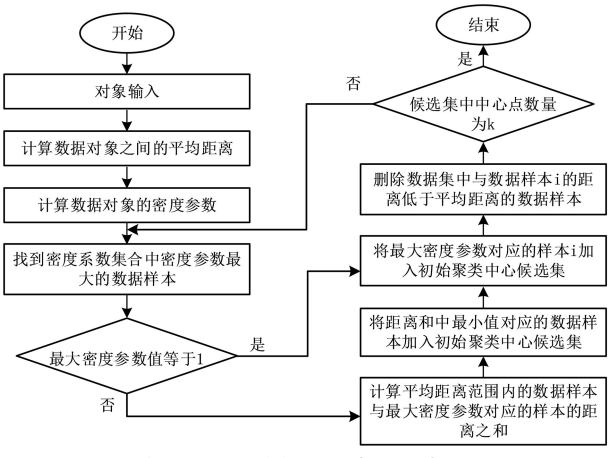

图 2 基于密度法的初始聚类中心点优化流程

数据样本集合 $\boldsymbol{S}=\{x_1,x_2,\cdots,x_n\}$ 中数据对象之

间的平均距离计算函数表示如下:

$$M(\boldsymbol{S}) = \frac{2}{n(n-1)} \times \sum_{i \neq j, i, j = 1}^{n} d(x_i, x_j) \tag{8}$$

式中:$M(\boldsymbol{S})$ 表示平均距离;$d(x_i, x_j)$ 表示数据对象 x_i 和数据对象 x_j 之间的距离。数据对象 x_i 的密度参数计算函数表示如下:

$$\begin{cases} \mathrm{den}(x_i, M) = \sum_{j=1}^{n} u(\mathrm{meandist} - d(x_i, x_j)) \\ u(x) = \begin{cases} 1, x > 0 \\ 0, \text{其他} \end{cases} \end{cases} \tag{9}$$

式中:meandist 表示平均距离。

计算数据集合 $\boldsymbol{S}$ 中的所有数据对象的密度参数,形成密度参数集合 $\boldsymbol{D} = \{ den(x_i, meandist), i \in (1, 2, \cdots, n) \}$。对集合 $\boldsymbol{D}$ 进行筛选,若数据集合 $\boldsymbol{D}$ 中密度参数最大的样本 i 的参数值等于 1,则将该数据对象加入初始聚类中心候选集,若密度参数最大值大于 1,并且样本点之间的距离低于平均距离,则最大密度参数所对应的平均距离范围内的所有点与距离之和的计算函数表示如下:

$$\mathrm{sum}(D_i) = \sum_{i=0}^{n_i} d(i, j) \tag{10}$$

将 $\mathrm{sum}(D_i)$ 的最小值所对应的数据对象加入初始聚类中心候选集中,删除密度参数集合中的数据对象 i,并删除数据集 $\boldsymbol{D}$ 中与数据对象 i 的距离低于平均距离的数据样本,反复迭代直至候选集中的聚类中心点数量为 k,这 k 个中心点即为算法的初始聚类中心点。通过密度法进行初始聚类中心点的选择,有效避免了传统 K-means 聚类算法的初始聚类中心点随机性较大的问题,提升初始聚类中心点位置选择的稳定性,减小初始聚类中心点位置对算法性能的不良影响。

2 实验与结果分析

为了验证改进 K-means 聚类算法的优化性和有效性,利用传统 K-means 聚类算法和改进 K-means 聚类算法对旅游电商平台数据进行处理,分别进行 20 次数据聚类分析实验,对 2 种算法的响应时间进行对比,对比结果如图 3 所示。

从图 3 中可以看出,传统 K-means 聚类算法的平均响应时间为 0.724 s,其中最大响应时间为 0.861 s,传统 K-means 聚类算法的运行速度较慢,系统响应时间较长。改进 K-means 聚类算法的平均响应时间为 0.498 s,其中最大响应时间为 0.647 s,经过优化改进,改进 K-means 聚类算法的平均响应响应时间缩短了 0.226 s,系统响应速度提升了 31.2%。通过随机梯度下降法和引入正则化项的方式对算法的损失函数进行优化,并利用自适应梯度优化算法自适应确定算法学习率,有效提升了 K-means 聚类算法的运行效率,加快算法收敛速度,从而缩短了改进 K-means 聚类算法的响应时间,具有较好的优化性,K-means 聚类算法的分析性能得到了明显提升。

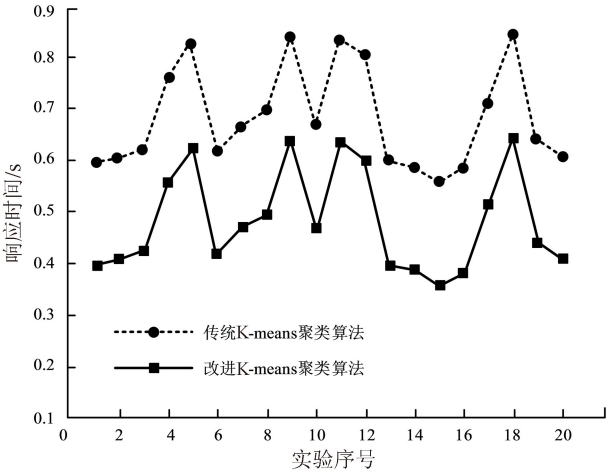

图 3　改进前后的 K-means 聚类算法性能对比

为了验证基于改进 K-means 聚类算法的旅游电商平台的实用性和可行性,采用线上实验的方式对改进 K-means 聚类算法的个性化推荐效果进行测试实验,并对浏览量等平台数据进行统计分析,基于改进 K-means 聚类算法的旅游电商平台的数据统计情况如表 1 所示。

表 1　基于改进 K-means 聚类算法的旅游电商平台数据

周次	平台浏览量/次	推荐产品浏览量/次	旅游产品购买量/次	推荐产品购买量/次	浏览行为的推荐点击率/%	购买行为的推荐成交率/%	推荐流量转化率/%
第 1 周	10 268	1 518	159	40	14.78	25.16	2.64
第 2 周	12 547	1 895	201	57	15.10	28.36	3.01
第 3 周	12 964	2 048	211	63	15.80	29.86	3.08
第 4 周	12 873	2 035	207	61	15.81	29.47	3.00
平均值	12 163						

从表 1 中可以看出,基于改进 K-means 聚类算法的旅游电商平台 4 周的平均每周平台浏览量为 12 163 次,其中经过改进 K-means 聚类算法个性化推荐的旅游产品的平均每周浏览量为 1 874 次,平台用户的浏览行为中推荐旅游产品的占比为 15.41%。旅游电商平台的每周平均产品购买量为 194.5 次,其中推荐旅游产品的每周平均购买量为 55.25 次,用户购买行为中推荐旅游产品的占比为 28.41%。结合改进 K-means 聚类算法的聚类分析

结果为用户进行个性化推荐,旅游电商平台的平均每周推荐流量转化率为29.48%,推荐流量有效转化为产品订单,基于改进K-means聚类算法的推荐流量质量较好,推荐流量向产品订单的转化率较高,能有效通过针对化的智能旅游产品推荐促成订单成交,提升电子商务平台的销售业绩。

利用改进K-means聚类算法、关联规则挖掘算法(Apriori)和基于用户的协同过滤算法(User-based CF)在旅游电商平台线上生产环境中进行在线实验,3种算法的旅游产品推荐成交情况如表2所示。

表2　3种算法的旅游产品推荐成交情况

算法	浏览行为推荐率/%	购买行为推荐率/%	推荐转化率/%
User--based CF算法	12.95	20.47	2.36
Apriori算法	13.44	23.63	2.58
改进K-means聚类算法	15.37	28.21	2.93

从表2中可以看出,基于改进K-means聚类算法的旅游电商平台的推荐流量转化率为2.93%,优于Apriori算法的2.58%和User-based CF算法的2.36%,推荐转化比例分别增加了0.35%和0.57%。在改进K-means聚类算法个性化推荐下,平台用户浏览行为中的推荐浏览率为28.21%,相较于Apriori算法和User-based CF算法分别提升了4.58%和7.74%,平台用户购买行为中的推荐旅游产品购买率为15.37%,相较于Apriori算法和User-based CF算法分别提升了1.93%和2.42%。利用改进K-means聚类算法构建旅游电商平台,为用户差异化地推荐符合其购买意向的旅游产品,能有效提高电商平台的旅游产品成交量,提升旅游企业的经济效益。

3 结论

随着电子商务行业的兴起,旅游产品销售方式发生变化,在线旅游产品预定方式成了一种趋势。为了提升旅游电子商务服务水平,基于K-means聚类算法构建旅游电子商务平台,并采用随机梯度下降算法、自适应梯度优化算法和密度法对K-means聚类算法进行优化改进,提升K-means聚类算法的收敛速度和运行性能。实验结果表明,改进K-means聚类算法的平均响应时间为0.498 s,系统响应速度相较于传统算法提升了31.2%,具有优化性。基于改进K-means聚类算法的旅游电子商务平台的推荐流量转化率为2.93%,平台用户浏览行为中的推荐浏览率为28.21%,平台用户购买行为中的推荐旅游产品购买率为15.37%,优于Apriori算法和User-based CF算法,能为平台用户提供个性化的旅游产品推荐,有效提升了旅游产品的购买成交量,具有较强的实用性和可行性,

参考文献:

[1] 杨单,刘启川.基于大数据的跨境电商平台个性化推荐策略优化[J].对外经贸实务,2020(11):33-36.
[2] 张磊,郭峰,侯小超.基于数据挖掘的电商搜索广告投放策略研究[J].工业工程,2019,22(1):69-78.
[3] 郭燕萍.电商客户数据挖掘中的模糊运算聚类算法分析[J].现代电子技术,2021,44(13):130-134.
[4] 阿荣,王丹琦.基于大数据分析的电子商务平台客户精准服务管理方法设计(英文)[J].机床与液压,2019,47(18):153-158.
[5] 陈婕.基于大数据技术的电商平台营销策略研究[J].福建茶叶,2019,41(11):18.
[6] 陈文行.大数据背景下电商平台信息分享策略研究[J].商业经济研究,2019(1):83-85.
[7] 李家华.基于大数据的人工智能跨境电商导购平台信息个性化推荐算法[J].科学技术与工程,2019,19(14):280-285.
[8] 孙华.大数据时代图书电子商务营销模式研究——以京东图书平台为例[J].出版广角,2020(16):56-58.
[9] 樊春美,朱建生.基于电商平台的恶意支付账户识别算法研究[J].计算机技术与发展,2020(6):114-118.
[10] 王二朋,倪郑宇.农产品线上消费者的偏好特征研究——基于"京东"销售苹果在线评论数据的分析[J].价格理论与实践,2020(2):120-123.
[11] 丁芙蓉,张功萱.基于CUDA并行化的K-Means聚类算法优化[J].计算机与数字工程,2019,47(7):1662-1666.
[12] 李艳.基于数据挖掘算法的移动电子商务群体用户访问控制模型[J].现代电子技术,2020(4):153-156.
[13] 申燕萍,顾苏杭,郑丽霞.基于云计算平台的仿生优化聚类数据挖掘算法[J].计算机科学,2019,46(11):247-250.
[14] 郑国凯,黄彩娥.基于大数据的智能商务分析平台开发和设计[J].现代电子技术,2020,43(5):163-166+170.
[15] 黄保华,程琪,袁鸿,等.基于距离与误差平方和的差分隐私K-means聚类算法[J].信息网络安全,2020(10):34-40.