杨进,杨孟,陈步前. 基于 K-means 和改进布谷鸟搜索算法的电影推荐[J]. 智能计算机与应用,2024,14(12):185-189. DOI:10.20169/j.issn.2095-2163.241227

基于 K-means 和改进布谷鸟搜索算法的电影推荐

杨 进,杨 孟,陈步前
(上海理工大学 理学院,上海 200093)

摘 要:针对电影推荐系统根据用户的喜好和大数据中的电影属性进行筛选时,因原始数据信息呈现海量化和稀疏化的特性,造成推荐准确率和用户满意度较低的问题,本文将 K-means 聚类和改进的布谷鸟搜索算法结合应用在数据集上对电影推荐系统做出改进。先对数据过滤后的数据集使用 K-means 算法进行聚类,再使用以锦标赛选择代替随机选择的布谷鸟搜索算法将一些项移动到更好的聚类中优化聚类结果,最后在构建的电影推荐系统上预测评分实现 Top-N 推荐。本文在 Movielens 数据集上进行实验,以平均绝对误差、均方根误差、准确率、召回率和 *F-Score* 为评价指标,与现有的算法相比,验证了所提方法的有效性。

关键词:推荐系统;布谷鸟搜索;K-means 聚类;数据过滤;用户喜好
中图分类号:TP391 **文献标志码**:A **文章编号**:2095-2163(2024)12-0185-05

Movie recommendation based on K-means and improved cuckoo search algorithm

YANG Jin, YANG Meng, CHEN Buqian
(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In order to solve the problem of low recommendation accuracy and user satisfaction due to the high quantification and sparseness of the original data information when the movie recommendation system is screened according to user preferences and movie attributes in big data, this paper combines K-means clustering and improved cuckoo search algorithm on the dataset to improve the movie recommendation system. Firstly, the dataset after data filtering is clustered using K-means algorithm, and then the cuckoo search algorithm that replaces random selection with tournament selection is used to move some items to better clusters to optimize clustering results. Finally, Top-N recommendation is achieved by predicting scores on the constructed movie recommendation system. In this paper, experiments are conducted on Movielens dataset, and the average absolute error, root-mean-square error, accuracy rate, recall rate and *F-Score* are taken as evaluation indicators. Compared with the existing algorithms, the effectiveness of the proposed method is verified.

Key words: recommendation system; cuckoo search; K-means clustering; data filtering; user preference

0 引 言

如今,推荐系统已成为电子商务不可分割的一部分,并被证明有助于为用户提供个性化的推荐。了解用户兴趣的一种方法是通过调查直接进行询问,但用户可能对填写表单不感兴趣,或者填写了不正确的信息。另一种更好的提取所需信息的方法是通过数据驱动的方式[1]。例如,协同过滤推荐系统通过其他用户过滤或评估项目。可根据过去有相似兴趣的用户的电影评分向另一个用户提供一些推荐。一般来说,推荐有 2 种不同的方式:根据用户的偏好提供一些推荐项目;预测用户尚未看过的电影项目的评分。

聚类作为一种无监督的工具,有助于根据相异性或相似性度量将数据集划分为不同类别。其中,K-means 聚类是一种高效简单的聚类算法,常被用于解决推荐问题。居晓媛等学者[2]提出一种融合标签文本的 K-means 聚类和矩阵分解的推荐算法,将 K-means 聚类应用到用户的潜在兴趣和项目的潜在特征提取中,提升了推荐算法的有效性。吴婷

基金项目:国家自然科学基金(12071293)。

作者简介:杨 进(1978—),女,博士,讲师,硕士生导师,主要研究方向:智能优化,图论与组合优化,Email:yangjin0903@163.com;杨 孟(1999—),男,硕士研究生,主要研究方向:推荐系统;陈步前(1998—),男,硕士研究生,主要研究方向:人工智能。

收稿日期:2023-07-01

婷等学者[3]提出一种基于 K-means 的改进协同过滤算法,将聚类算法和协同过滤算法进行结合,并对聚类算法和相似度计算进行相应的改进,提高了推荐的效果。但是 K-means 聚类算法受初始聚类中心影响过大且容易陷入局部最优。

在处理因原始数据信息的海量性和稀疏性造成推荐准确率和用户满意度较低的问题时,单一使用 K-means 聚类算法容易陷入局部最优。元启发式优化算法以其在短时内解决大规模优化问题的潜力而被广泛使用,且具有一定的全局搜索能力。许多研究人员将聚类和元启发式优化算法结合,优化聚类效果,从而提高推荐质量。李艳娟等学者[4]提出了一种基于改进蜂群 K-means 聚类模型的协同过滤推荐算法,采用最大最小距离积邻域均值法初始化聚类中心,并提出了新的适应度函数,以提高推荐质量和推荐效率。胡安明[5]提出一种基于自适应布谷鸟聚类搜索的改进推荐系统算法,首先对推荐数据进行聚类处理,然后利用布谷鸟算法较强的全局搜索能力,提升推荐系统的准确度。这些利用元启发式算法优化聚类结果的方法可以增强典型的基于内容和协同过滤推荐系统的预测能力。其中,布谷鸟搜索是一种结合布谷鸟巢寄生性和 Lévy 飞行模式的新兴的元启发式算法,并且布谷鸟搜索和布谷鸟搜索的自适应版本已经显示出良好的基准测试结果[6],但布谷鸟搜索算法使用随机选择方法,可能会忽略其他更好的解决方案。因此,本文提出一种基于 K-means 和改进布谷鸟搜索算法的电影推荐方法。在本文的方法中,首先,对数据集进行过滤;然后,使用 K-means 对数据进行聚类,并利用"肘方法"确定聚类的最佳簇数;接着,结合改进的布谷鸟搜索算法优化聚类结果,其中布谷鸟搜索算法的改进是将随机选择法替换为锦标赛选择法;最后,在构建的电影推荐系统上预测评分实现 Top - n 推荐。通过数据过滤和改进的布谷鸟搜索,可以找到让用户更满意的推荐。文中使用平均绝对误差、均方根误差、准确性、召回率和 F - *Score* 作为评价指标,验证了所提方法的有效性。

1 相关工作

1.1 推荐系统

由于信息过载,有时用户无法在网站上找到相关和准确的信息。为了解决这些问题,推荐系统被用于筛选和排序数据,向用户推荐关键信息。协同过滤算法是推荐技术领域最先出现的算法之一,并且也是应用得最广泛的推荐算法。协同过滤的原理是先给目标用户找出其相似用户,再根据相似用户的偏好来挖掘出目标用户的偏好[7]。电影推荐系统一般是基于用户的协同过滤推荐,电影推荐系统会通过用户行为(打分或相似度)的异同做出不同推荐。由于用户数据通常存在稀疏性、可扩展性、同义性和冷启动性,导致推荐质量下降,许多电影推荐使用协同过滤和聚类来进行改进[8-9]。

1.2 聚类

聚类算法属于无监督学习方法,可以准确分析大量数据。聚类先将项目分成相似的集合(称为簇),同一簇内的项目之间的差异最小,再构建模型。当一个新用户出现时,模型计算其与当前簇的相似度,并将其分配到与该用户最相似的簇中。在众多聚类算法中,本文选择 K-means 聚类算法,因为原理比较简单,收敛速度快,算法的可解释性较强,并且需要调参的参数主要是簇数 K[10]。K-means 算法步骤具体如下。

算法 1 K-means 聚类

输入 样本集 D,聚类的簇数 K

输出 簇划分 C

1. 数据准备;

2. 对于未聚类数据集,首先从数据集 D 中随机选择 K 个样本作为初始的 K 个质心;

3. 求出每个样本到质心的距离,按照距离自身最近的质心进行第一次聚类;

4. 依据上次聚类结果,求出新的质心(新簇内点 x 和 y 的平均值);

5. 反复迭代,直到中心点的变化满足收敛条件(变化很小或几乎不变化),最终得到聚类结果。

1.3 布谷鸟搜索

布谷鸟搜索算法是受自然启发的元启发式算法,采用了一种特定的搜索方法,为每个项或用户选择更好的聚类[11]。布谷鸟的产卵行为假设有 3 个理想条件:每一只布谷鸟都会生产一个蛋,并被随机扔在一个鸟巢里;具有优质蛋的最佳的巢被传到下一代;鸟巢的数量是已知的,鸟窝主人有一定的概率发现外来的布谷鸟蛋。如果鸟窝主人认出这个外来的蛋,那么就会把蛋扔掉,或者放弃整个鸟巢。布谷鸟搜索算法的搜索分为全局搜索和局部搜索,分别占总搜索时间的 3/4 和 1/4,这就使得可以在全局范围内进行更有效的探索,从而提高寻优效率[12]。与采用标准高斯过程的算法相比,由于 Lévy 飞行机制存在着无穷的方差和均值,使得布谷鸟搜索能够

更高效地探索搜索空间。布谷鸟搜索与灰狼优化[13]、禁忌搜索[14]、最近邻搜索[15]等算法结合使用可以解决不同的问题，其中包括推荐系统。文献[16]提出了一种传统布谷鸟搜索和简单 K-means 的电影推荐。在文献[17]中，改进布谷鸟搜索可以改进局部搜索，并且布谷鸟巢可以交换信息。在文献[18]中，使用了改进的布谷鸟搜索和改进的聚类，用高斯指数函数代替 Lévy 飞行函数，并用模糊 C 均值聚类代替聚类算法。布谷鸟搜索算法步骤描述如下。

算法 2　布谷鸟搜索

输入　鸟窝规模 N，维度 D，发现概率 pa，鸟巢界值，最大迭代次数 Max$N$

输出　最优鸟巢位置 X_{best}^0，最优解 $f_{\min}$

1. 确定目标函数 $f(x)$，$X=(x_1,\cdots,x_d)^{\mathrm{T}}$ 初始化群体，随机产生 n 个鸟巢的初始位置 $X_i(i=1,2,\cdots,n)$；

2. 采用 Lévy 飞行机制更新当代鸟巢的位置；将当代鸟巢与上一代鸟巢位置 $P_{t-1}=[X_1^{t-1},X_2^{t-1},\cdots,X_n^{t-1}]^{\mathrm{T}}$ 进行对比，用适应度值较好的鸟巢位置代替适应度值较差的鸟巢位置：$G_t=[X_1^t,X_2^t,\cdots,X_n^t]^{\mathrm{T}}$；

3. 用随机数 R 作为鸟巢主人发现布谷鸟蛋的概率，将其与鸟被淘汰的概率 pa 进行比较。若 $R>pa$，则随机改变 G_t 中的鸟窝位置，得到一组新的鸟窝位置。再更新鸟窝位置，得到一组较好的鸟窝位置：$p_t=[X_1^t,X_2^t,\cdots,X_n^t]^{\mathrm{T}}$。更新最优鸟窝位置 X_{best}^t 和最优解 $f_{\min}^t$；

4. 判断算法是否满足设置的最大迭代次数：若满足，结束搜索过程，输出全局最优值 $f_{\min}$，否则，重复步骤 2 进行迭代寻优。

2　基于 K-means 和改进的布谷鸟搜索算法的电影推荐

2.1　改进的布谷鸟搜索算法

在传统的布谷鸟搜索中，先由 Lévy 飞行产生首次的解决方案，再随机选择另一种解决方案与之比较适用性。虽然传统的布谷鸟搜索通过与其它元启发式算法的比较证明了其效率，但这种算法随机选择解的方式可能会降低结果的多样性，忽略适应度值较大的解，影响数据聚类的效果。在本文中，使用锦标赛选择法代替随机选择法。锦标赛选择法的步骤为：

(1)设定每次选取的个体数目 M(选取2个个体为二元锦标赛选择)；

(2)按照相同的概率从种群中随机选取 M 个个体，根据每个个体的适应度值大小，选择其中适应度值最大的个体进入下一代种群；

(3)重复步骤 2(次数为种群的大小)，直到新的种群规模达到原先的种群规模。

本文通过锦标赛选择从种群中选择一个解决方案，再将其与通过 Lévy 飞行选择的另一个解决方案进行比较。于是，一些比赛将在可能的候选人之间进行，获胜者将取代当前的解决方案，能够产生更可靠的数据聚类。在本文的电影推荐系统中，用户被视为布谷鸟蛋，并通过 K-means 聚类在不同簇中。如果布谷鸟蛋 x_n 的适应度值超过簇中当前用户的特定百分比，将替换该用户。当将现有的蛋换成更好的类似的蛋时，就不会被发现且被留在巢中。当那些布谷鸟的蛋孵化时，则会尝试从巢中随机扔出一个蛋。图 1 为改进的布谷鸟搜索流程。

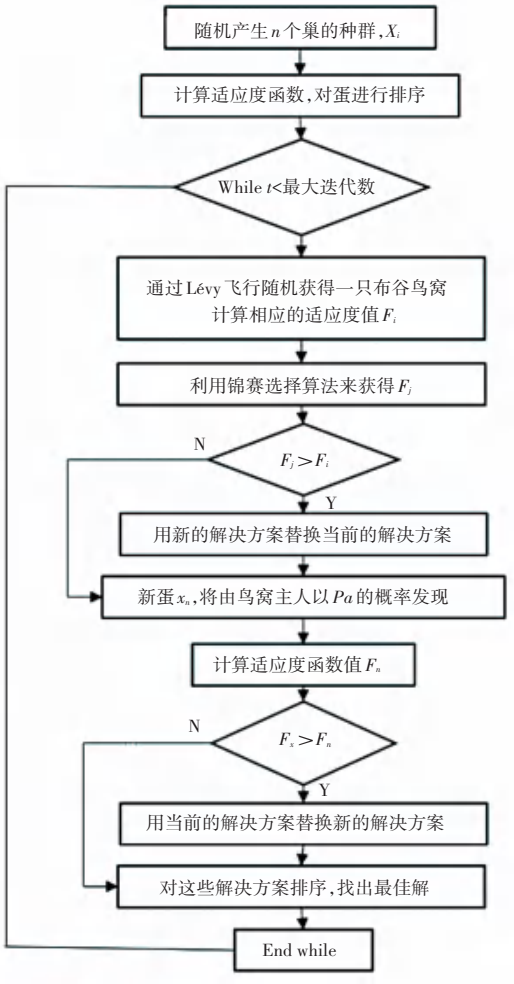

图 1　改进的布谷鸟搜索算法流程图

Fig. 1　Flowchart of improved cuckoo search process

2.2 方法描述

本文的方法是结合聚类和元启发式算法来优化聚类效果,从而提高推荐质量。在本文的方法中,先采用 K-means 算法对数据进行聚类,再结合改进的布谷鸟搜索算法优化聚类结果,其中布谷鸟搜索算法的改进是将随机选择法替换为锦标赛选择法。步骤如下。

(1)使用 K-means 将 MovieLens 数据集的用户划分成不同的簇:首先随机选择聚类的质心,然后计算每个用户到每个质心的距离,最后把所有用户分配到最近的簇。在第一组分配之后,计算每个用户与新的聚类质心的距离,如欧几里得、余弦等,根据距离将其重新分配到最近的聚类质心。这个迭代过程将继续下去,直到不再发生重新分配或达到最大迭代数;

(2)将改进的布谷鸟搜索算法应用于聚类的结果:使用改进的布谷鸟搜索算法为前面的每个聚类计算适应度函数,用来表示解的质量,并计算到每个质心的距离将用户重新分配到最近的簇,最后再次应用 K-means 直到质心固定。

3 实验结果

3.1 实验数据集和实验环境

本文在 MovieLens 1M 数据集上进行了应用和测试,包含来自 6 040 名 MovieLens 用户的 3 952 部电影的 1 000 209 个评分。仿真中过滤掉所有评分在 4 分以上的电影,向评分高的用户推荐新电影,以获得更高质量的预测。因为用户通常愿意观看评分较高的电影,而评分为 1 分(满分 5 分)的电影可能缺少吸引力。

本文代码是用 Python 3.9 编写的,运行在 Intel corei5 CPU 和 Nvidia GeForce GTX 1650Ti 上,开发工具为 PyCharm IDE 和 Jupyter Notebook。

3.2 评价指标

(1)平均绝对误差 (*MAE*):是理解推荐系统行为的最常用指标。

$$MAE = \frac{\sum | P_{ij} - R_{ij} |}{n} \tag{1}$$

其中,P_{ij} 表示用户的预测值,R_{ij} 表示先前已知的电影评分。

(2)均方根误差 (*RMAE*):通过对所有的真实评分和预测评分取差值来衡量推荐的准确度。

$$RMAE = \sqrt{\frac{\sum (P_{ij} - R_{ij})^2}{n}} \tag{2}$$

(3)准确率:是目标用户检索到的相关项目的数量与推荐总数的比值,反映测量值与实际真实值的接近程度。

$$precision = \frac{t_p}{t_p + f_p} \tag{3}$$

其中,t_p 表示分类正确,把原本属于正类的样本分成正类;f_p 表示分类错误,把原本属于负类的错分成正类。

(4)召回率:表明了算法的灵敏度,旨在找到实际为正的样本中多少被预测为正。

$$recall = \frac{t_p}{t_p + f_n} \tag{4}$$

其中,f_n 表示分类错误,把原本属于正类的错分成了负类。

(5) $F - Score$:是准确率和召回率的加权平均值。

$$F - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5}$$

3.3 结果分析

本文使用 Flask 框架制作了一个在线推荐系统,该系统使用经过训练的模型,向目标用户推荐评分前 10 的电影。

在实验的第一步,用户评分矩阵被创建并用于 K-means 聚类。在如 K-means 聚类这样的无监督聚类方法中,选择 K 的最佳数量是一个问题。本文使用“肘方法”将数据变化绘制成簇数的函数,并选择曲线的弯头作为要使用的最佳簇数。为此,需要找到组内平方和最小的簇内节点平方偏差之和 (*WCSS*)。*WCSS* 表示每个样本点和簇内质心的距离偏差之和。簇划分得越多,每个簇的聚合程度就越高,*WCSS* 组内平方和越小。

本文在聚类过程中将少于 25 个用户的簇组合在一起,并将其全部放在一个未分类的簇中,可以减少聚类的簇数。实验结果如图 2 所示。在图 2 中,随着 K 值的增大。即质心的增多,整体的 *WCSS* 是逐渐减小的(因为每个点能找到与其距离更近的质心的概率变大了),所以可以通过不断增大 K,来观察整体的性能。随着簇的数量的增加,每个簇的用户数量也会更少,每个用户可能会被分配到更合适的簇中,每个簇中聚类的相似性更大,使得预测误差较低。簇的最优数量是不同簇的差异足够大,内部又足够相似。从 $K = 15$ 开始,*WCSS* 不再出现明显下降,说明 $K = 15$ 是最优选择。

在实验的第二步,将改进的布谷鸟搜索算法应用于聚类的结果。本文先将过滤后的数据集分别应用于没有任何优化算法的方法:K-means 以及 PCA 和 K-means。接着,在一个实验中加入传统的布谷

鸟搜索算法，在另一个实验中加入修改的布谷鸟搜索算法，来验证修改后算法对 K-means 聚类结果的改善程度。这里对本文提出的模型进行了 10 次交叉验证，平均结果见表 1，并在表 2 中比较了这些算法的训练时间和测试时间。可以看出，与其它算法相比，数据集在经过 K-means 聚类和改进的布谷鸟搜索算法处理后，推荐系统产生的预测结果在各项评价指标上表现最好，算法的运行时间也是最快的。以上结果表明，本文方法可以得到更可靠的聚类结果，从而提高推荐的质量和效率。

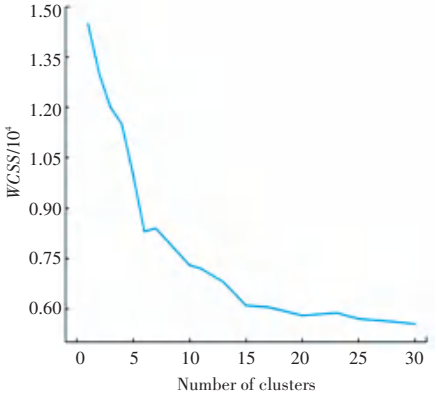

图 2　实验结果

Fig. 2　Experimental results

表 1　不同算法的结果对比

Table 1　Comparison of the results of different algorithms　%

算法	*MAE*	*RMAE*	*Precision*	*Recall*	*F - Score*
K-means	85	93	31	15	20
PCA 和 K-means	77	88	40	18	25
传统的布谷鸟搜索和 K-means	66	82	55	36	44
改进的布谷鸟搜索和 K-means	**61**	**77**	**67**	**41**	**51**

表 2　算法运行时间对比

Table 2　Comparison of algorithms runtime　s

算法	训练时间	测试时间
K-means	26	16
PCA 和 K-means	19	12
传统的布谷鸟搜索和 K-means	46	15
改进的布谷鸟搜索和 K-means	**24**	**11**

4　结束语

本文研究在 MovieLens 数据集上应用 K-means 聚类和改进的布谷鸟搜索，提出了一种利用该数据集改进原有电影推荐系统的新方法，使聚类性能比以前更好、更快。在此基础上，使用平均绝对误差、均方根误差、准确性、召回率和 *F - Score* 对结果进行了评估，验证了本文方法的有效性。在未来的工作中，可以使用其他自然启发的元启发式算法来代替改进的布谷鸟搜索，也可以使用其他适应度函数来代替锦标赛算法。

参考文献

[1] 让冉，邢林林，张龙波，等. 面向新领域的推荐系统综述[J]. 智能计算机与应用，2023，13(5)：1-8.
[2] 居晓媛，汪明艳. 融合标签文本的 K-Means 聚类和矩阵分解算法[J]. 软件工程，2023，26(6)：30-35.
[3] 吴婷婷，李孝忠，刘徐洲. 基于 K-Means 的改进协同过滤算法[J]. 天津科技大学学报，2021，36(6)：44-48.
[4] 李艳娟，牛梦婷，李林辉. 基于蜂群 K-means 聚类模型的协同过滤推荐算法[J]. 计算机工程与科学，2019，41(6)：1101-1109.
[5] 胡安明. 基于自适应布谷鸟聚类搜索的推荐系统算法的研究[J]. 电脑知识与技术，2022，18(6)：87-88.
[6] SAKGOTRA R，SINGH U，SAHA S，et al. Self adaptive cuckoo search：Analysis and experimentation [J]. Journal of Swarm and Evolutionary Computation，2021，60：100751.
[7] 司品印，齐亚莉，王晶. 基于协同过滤算法的个性化电影推荐系统的实现[J]. 北京印刷学院学报，2023，31(6)：45-52.
[8] 王伟峰，李澍源. 基于混合聚类优化协同过滤的 Web 服务推荐[J]. 信息技术，2022，46(11)：44-48.
[9] 苏凯，张萱，付静. 基于项目属性聚类及相似度优化的协同过滤算法[J]. 海军工程大学学报，2022，34(2)：20-26.
[10] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.
[11] 包天悦，高剑飞，王永咏，等. 基于动态布谷鸟算法的无线传感网能量优化方法研究[J]. 自动化与仪器仪表，2022(9)：22-25.
[12] WALTON S，HASSAN O，MORGAN K，et al. Modified cuckoo search：A new gradient free optimization algorithm [J]. Chaos Solitons Fractals，2011(44)：710-718.
[13] DEB S，CHATUANRAMTHRNGKAKA B，DATTA S，et al. Congestion management by generator real power rescheduling using hybrid grey wolf optimizer and cuckoo search algorithm [C]// 2021 1st International Conference on Power Electronics and Energy (ICPEE). Piscataway，NJ：IEEE，2021：1-5.
[14] TERKI A，BOUBERTAKH H. A new hybrid binary-real coded Cuckoo Search and Tabu Search algorithm for solving the unit-commitment problem [J]. International Journal of Energy Optimization and Engineering，2021，10(2)：104-109.
[15] GARCÍA J，MAUREIRA C. A KNN quantum cuckoo search algorithm applied to the multidimensional knapsack problem [J]. International Journal of Applied Soft Computing，2021，102：107077.
[16] SINGH S，SOLANKI S. A movie recommender system using modified cuckoo search [M]// SRIDHAR V，PADMA M，RAO K. Emerging Research in Electronics，Computer Science and Technology. Lecture Notes in Electrical Engineering. Cham：Springer，2019：471-482.
[17] SALGOTRA R，SINGH U，SAHA S，et al. Self adaptive cuckoo search：Analysis and experimentation [J]. Journal of Swarm and Evolutionary Computation，2021，60：100751.
[18] SELVI C，SIVASANKAR E. A novel optimization algorithm for recommender system using modified fuzzy c-means clustering approach [J]. Soft Computing，2019，23：1901-1916.