融合协同过滤的 CatBoost 推荐算法

唐 震,黄 刚,华雯丽
(南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210023)

摘 要:在推荐系统中,针对推荐准确度问题,提出了一种融合协同过滤和 CatBoost 的混合推荐算法(UCF-CB)。在协同过滤模块中对用户相似度计算公式进行改进,加入时间衰减因子以及热门物品惩罚项,利用改进后的协同过滤算法对用户项目评分矩阵进行评分预测,得到用户对物品的一次评分。对协同过滤一次评分进行降序排序,选取评分最高的前 k 项物品,形成召回集。对原始数据集进行预处理,挖掘潜在特征增加特征维度,利用 CatBoost 算法对用户和项目特征进行训练,对召回集数据进行预测,得到二次评分预测。对于没有评分记录的新用户,利用训练好的 CatBoost 算法可以直接进行评分预测,在一定程度上解决了推荐系统冷启动的问题。将协同过滤一次评分以及 CatBoost 二次预测评分进行加权融合得到更为精确的推荐结果。在 movieslens(ml-1m)数据集上的实验结果表明,该算法可以获得较高的准确度。

关键词:CatBoost;协同过滤;准确性;推荐系统;混合模型
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2021)09-0036-07
doi:10.3969/j.issn.1673-629X.2021.09.007

CatBoost Recommendation Algorithm with Collaborative Filtering

TANG Zhen, HUANG Gang, HUA Wen-li
(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: In the recommendation system, a hybrid recommendation algorithm (UCF-CB) which combines collaborative filtering and CatBoost is proposed to improve the accuracy of recommendation algorithm. In the collaborative filtering module, the user similarity calculation formula is improved by adding time attenuation factor and hot item penalty factor. The improved collaborative filtering algorithm is used to predict the user's item rating matrix, then the user's first prediction score is obtained. The first score of collaborative filtering is sorted in descending order, and the top k items with the highest score are selected to form the recall set. After preprocessing the original data set, mining potential features and increasing feature dimensions, CatBoost algorithm is used to train the features of users and items, the trained model is used to predict the recall set data, and the second score prediction is obtained. For the new users without scoring records, the trained CatBoost algorithm can directly predict the score, which solves the problem of cold start of recommendation system to some extent. The first prediction score of collaborative filtering and the second prediction score of CatBoost are weighted to get more accurate recommendation results. Experiments on movieslens (ml-1m) data set show that the proposed algorithm can achieve high accuracy.

Key words: CatBoost; collaborative filtering; accuracy; recommendation system; hybrid model

0 引 言

随着信息技术和互联网环境的不断发展,网络用户的不断增加,人们进入了大数据的信息时代。电子商务市场和社交媒体的逐渐壮大是信息和数据急速增长的主要原因,网络数据的快速增长在丰富互联网生活的同时产生大量无用的数据,这些冗余信息给人们的生活带来了不便。用户对信息的反应速度无法跟上信息的传输及更新速度,用户会对信息的选择产生偏差,这导致了信息利用率的下降,这种类型的问题称为信息过载[1]。解决信息过载的方法一般有两种,其中一种为搜索引擎,利用用户指定的关键字,通过关键词检索被动地过滤出用户感兴趣的信息,速度快但并不能突出用户的个体需求。另一个便是推荐系统,推荐系统通过提供信息过滤机制来帮助用户处理信息过载问题,能够更好地满足用户个性化需求。

推荐系统大致可分为三类:基于内容的推荐[2]、协同过滤[3]推荐方法和混合推荐方法[4]。基于内容的推荐是根据物品的相关信息,用户相关信息以及用户

收稿日期:2020-10-20 **修回日期**:2021-02-25
基金项目:江苏省教育基金资助项目(17JS010);中国电信公司江苏分公司基金资助项目(DGJ02)
作者简介:唐 震(1996-),男,硕士,研究方向为推荐系统、机器学习;黄 刚,教授,研究方向为数据挖掘。

相关行为构建模型找出用户喜爱的物品。协同过滤通过用户反馈信息进行推荐,有两种常用类型:基于用户的协同过滤(UCF)和基于项目的协同过滤(ICF)。混合推荐方法将不同的推荐算法过程或结果进行结合得出最终的推荐结果,组合后能够弥补单个推荐算法的缺点。

协同过滤算法是目前研究最多、应用最广泛的经典推荐算法。该算法实现简单,适用性强,推荐效果较好。但随着互联网数据的几何式增长,协同过滤推荐系统出现了很多问题,由于协同过滤需要利用用户评分矩阵,所以当项目用户过多时,部分评分缺失,计算相似度困难。面对新用户,没有足够的历史数据,无法计算目标用户的相似用户群,给推荐带来了障碍,这是推荐系统中经典的冷启动问题[5]。并且当数据量过大,推荐系统难以避免的会出现推荐速度慢,推荐准确度低的问题。

解决冷启动问题一般通过利用用户的注册信息,用户上下文信息,基于热门数据的推荐等方法。Koren Y 等人[6]提出了基于矩阵分解的系统过滤算法,矩阵分解的一个优点是它允许合并额外的信息。当没有明确的反馈时,推荐系统可以使用隐式反馈来推断用户偏好,隐式反馈通过观察用户行为来间接反映用户的偏好。张峰等人[7]提出了使用 BP 神经网络缓解协同过滤推荐算法的冷启动问题,根据用户评分向量交集大小选择候选最近邻居集,采用 BP 神经网络预测用户对项的评分,减小候选最近邻数据集的稀疏性,解决了冷启动问题。

推荐系统准确度问题通过混合推荐算法、先进的评分预测算法等加以解决。Cheng-Che Lu 和 VS-Tseng[8]提出基于内容、基于情感的协同过滤推荐算法,该方法通过对用户选择的反馈,适应用户兴趣的变化,进而推荐出自己喜欢的、更有趣的物品,并进行即时连续推荐。

文中提出了一种融合协同过滤算法和 CatBoost[9]的推荐算法,该算法旨在提高推荐系统模型的准确度,实现个性化推荐,提高用户的满意度,并在一定程度上缓解推荐系统冷启动问题。算法思路:通过协同过滤算法构建用户相似度矩阵,根据预测评分公式计算出用户对未评分物品的预测评分,由此得到候选集,再利用 CatBoost 算法对整体数据集进行训练,对候选集进行预测评分,最后将两个预测评分进行加权融合,得到最终的预测评分从而实现对用户的推荐。由于利用 CatBoost 算法进行数据集训练利用到的特征较多,可以更好地实现用户个性化推荐,与协同过滤算法进行加权融合后得到的结果更加准确。对于新用户或者历史交互记录缺少的用户,直接用 CatBoost 算法进行预测并推荐,可以有效缓解用户的冷启动问题。

1　相关工作

1.1　协同过滤算法

协同过滤算法(collaborative filtering recommendation)仍然是目前最为流行、使用最为广泛的推荐算法。该算法的整体推荐效果在很多场景中不亚于新研究出来的其他推荐算法,并且相比较其他算法,协同过滤推荐算法的性能最为稳定。协同过滤算法通过系统提供的用户项反馈来产生推荐,目的是在反馈信息中查找推荐信息,不需要关于用户或项目的额外数据,可以作为推荐系统候选生成器。

基于用户的协作过滤通过分析用户的历史行为数据,然后根据不同用户对相同物品的评分或偏好程度来评测用户之间的相似性,对有相同偏好的用户进行物品推荐。基于项目的协同过滤通过对不同物品的评分来预测项目之间的相似性,再根据用户历史行为数据,得到用户喜欢的物品,通过项目相似性矩阵找到相似度最高的物品,从而推荐给用户。

基于用户的协同过滤和基于项目的协同过滤的核心思想是在整个数据空间寻找用户或项目的前 k 个最近邻,都需要进行相似度计算。相似度计算的常用方法有余弦相似度、皮尔森(Pearson)相关系数法、欧几里得距离法、杰卡德相似系数法等。

余弦相似度计算公式如下:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I}R_{u,i}\bullet R_{v,i}}{\sqrt{\sum_{i\in I}R_{u,i}^2}\bullet\sqrt{\sum_{i\in I}R_{v,i}^2}} \tag{1}$$

其中,I 为所有项目的集合,i 为项目集合中的单个项目,$R_{u,i}$ 为用户 u 对项目 i 的真实评分,$R_{v,i}$ 为用户 v 对项目 i 的真实评分。

皮尔森相关系数计算公式为:

$$\mathrm{sim}(u,v)=\frac{\sum_{u\in U}(R_{u,i}-\bar{R}_i)\bullet(R_{u,v}-\bar{R}_v)}{\sqrt{\sum_{u\in U}(R_{u,i}-\bar{R}_v)^2}\bullet\sqrt{\sum_{u\in U}(R_{u,v}-\bar{R}_v)^2}} \tag{2}$$

其中,U 为所有对项目评分的用户集合,$\bar{R}_i$,$\bar{R}_v$ 是项目 i 和 v 的平均得分,最终的相似度得分在 $[-1,1]$ 之间。

欧几里得距离公式为:

$$\mathrm{sim}(u,v)=\frac{1}{1+\sqrt{\sum_{i\in I}(R_{u,i}-R_{v,i})^2}} \tag{3}$$

杰卡德系数为:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I}R_{u,i}\bullet R_{v,i}}{\sum_{i\in I}R_{u,i}^2+\sum_{i\in I}R_{v,i}^2-\sum_{i\in I}R_{u,i}\bullet R_{v,i}} \tag{4}$$

通过相似度,可以预测得到用户对某一项目的评分,从而将评分高的项目推荐给用户。定义 U 为与目标用户的相似用户集,$\bar{R}_u$ 为用户 u 对已评分项目的平均评分,预测评分为:

$$R(u,i)=\bar{R}_u+\frac{\sum_{v\in U}\mathrm{sim}(u,v)\bullet(R_{v,i}-\bar{R}_v)}{\sum_{v\in U}\mathrm{sim}(u,v)} \tag{5}$$

在协同过滤算法中加入热门商品惩罚项和时间衰减因子可以提高个性化推荐的准确度。惩罚热门物品的原因是如果一个物品过于热门,会有很多用户对其进行评分,但这并不能说明这些用户有着相同的兴趣,所以对热门物品增加一个惩罚项,减少热门物品对用户相似度的影响。

热门商品惩罚项计算公式为:

$$\frac{1}{\lg(1+N(i))} \tag{6}$$

其中,$N(i)$ 表示对物品 i 进行过评分的所有用户。

由于用户最近的行为更能表达用户的当前兴趣,所以在计算用户相似度时可以增加时间衰减函数。时间衰减函数为:

$$f(|t_{ui}-t_{vi}|)=\frac{1}{1+\alpha|t_{ui}-t_{vi}|} \tag{7}$$

其中,t_{ui} 表示用户 u 对物品 i 产生行为的时间,t_{vi} 表示用户 v 对物品 i 产生行为的时间。通过对相似度计算方法的改进,得到优化后的余弦相似度公式:

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I}\frac{1}{\lg(1+N(i))}\bullet f(|t_{ui}-t_{vi}|)\bullet R_{u,i}\bullet R_{v,i}}{\sqrt{\sum_{i\in I}R_{u,i}^2}\bullet\sqrt{\sum_{i\in I}R_{v,i}^2}} \tag{8}$$

1.2 CatBoost

CatBoost 全称为 Gradient Boosting(梯度提升)+ Categorical Features(类别型特征),是一种对决策树进行梯度增强的算法,属于集成学习算法的一种。它由 Yandex 公司的研究人员和工程师开发,用于 Yandex 和其他公司的搜索、推荐系统、个人助理、自动驾驶汽车、天气预报和许多其他任务[10]。

Gradient Boosting 方法是 Boosting 方法的一种,Boosting 模型是通过最小化损失函数得到最优模型,是一个迭代的过程,每一次新的训练的目的是改进前一次训练的效果。Gradient Boosting 的主要思想是每次都在前一次模型损失函数的负梯度方向建立新的模型使得损失函数能够不断下降[11]。Gradient Boosting 方法适用于异质化数据,梯度提升方法比神经网络的入门门槛更低,使用起来也更简单。近年来,不少学者尝试将集成学习算法运用到推荐系统中,崔岩等[12]提出融合协同过滤和 XGBoost 的推荐算法,提高了推荐的准确性;李智彬[13]提出融合 SVD 与 LightGBM 的音乐推荐算法,解决推荐系统冷启动问题。作为最新的集成学习算法由于对 CatBoost 算法的相关研究较少,还没有学者将其运用在推荐算法中。

CatBoost 是一种能够很好地处理类别型特征的梯度提升算法库。根据官方测评结果[14],CatBoost 在准确率方面比同类型的 XGBoost 以及 LightGBM 表现更加优秀,该测评结果是在部分数据集上进行的实验,在大多数实验对比中,CatBoost 都有着较为不错的训练速度与准确率。

CatBoost 相比其他梯度提升算法具有两大优势:第一,不需要人为地处理类别型特征,CatBoost 算法可以直接使用类别特征进行模型训练,CatBoost 使用独特的方法处理类别特征[15]。首先对样本进行随机排序,然后针对类别型特征中的某个取值,每个样本的该特征转为数值型时都是基于排在该样本之前的类别标签取均值,同时加入了优先级和优先级的权重系数。并且可以将类别特征进行组合,利用特征之间的联系,这极大地丰富了特征维度。定义编码值的公式为:

$$\hat{x}_k^i=\frac{\sum_{x_j\in D_k}1[x_j^i=x_k^i]\bullet y_j+ap}{\sum_{x_j\in D_k}1[x_j^i=x_k^i]+a} \tag{9}$$

其中,x_k^i 为样本 k 的第 i 个特征,D 为所有可用于模型训练的数据集,D_k 为 D 中的子集,y_j 为样本 j 的特征值,a 为参数,p 为先验值,添加先验项是一个普遍做法,针对类别数较少的特征,它可以减少噪音数据,通常设置为数据集中标签的平均值,1 为数学公式 indicator function,当后面括号的内容为真时取值 1,否则取值 0。

类别特征的任何组合都可以视为新特征。例如,假设任务是电影推荐,有两个特征:用户 ID 和电影类型,某些用户喜欢战争类的电影。在将用户 ID 和电影类型转换为数字特征时,会丢失此信息。将用户 ID 与电影类型两种特征进行组合解决了这个问题,并提供了一个新的特征。

在数据集中,特征组合数与特征数为指数关系,对于特征较多的数据集不可能在算法中考虑所有组合,这样会增加计算量。CatBoost 在决策树的新一轮拆分时,以贪婪的方式对特征进行组合[16]:对于树中的第一个拆分不考虑组合。在下一个分割节点选择时,CatBoost 将所有组合特征和分类特征与数据集中的所有分类特征组合在一起。组合值会动态转换为数字,通过计算它的 TS(target statistics)值作为新的特征值参与树模型构建。CatBoost 通过以下方式生成数字和类别特征的组合:在决策树中,所有的拆分都被作为具有两个值的类别特征,采用与类别特征相同的组合方

式进行组合使用。

CatBoost 相对于其他梯度提升算法的第二个优势:在选择生成树结构时,计算叶子节点的算法可以避免过拟合。传统的 GBDT 算法存在由于梯度估计偏差引起的过拟合问题,预测偏差是由一种特殊类型的目标泄漏引起的。CatBoost 提出使用 Ordered boosting[16] 的方法来解决预测偏差从而得到梯度步长的无偏估计。Ordered boosting 算法首先会生成一个长度为 n 的序列,对每个样本 x_i 训练出一个单独的模型 M_i ,使得 M_i 是仅利用了序列中的前 i 个样本,不包含 x_i 的训练集得到的训练模型。利用 M_{j-1} 训练模型得到第 j 个样本的梯度估计。

CatBoost 同时具有 CPU 和 GPU 实现。GPU 的实现允许更快的训练,同时还具有快速的 CPU 评分实现。对于数值密集型特征的训练,最重要的就是找到最佳分割,这是 GBDT 算法最主要的计算负担。CatBoost 使用对称决策树作为基础学习者,并将特征离散化为固定数量的箱(bins),以减少内存使用量,在训练模型时可以设置最大箱数[17]。

2　融合协同过滤与 CatBoost 算法的推荐模型

2.1　算法思想

本模型融合 CatBoost 与协同过滤的算法为用户进行推荐,形成新的算法模型 UCF-CB。由于传统的协同过滤算法根据用户评分信息计算用户对评分物品的预测评分从而进行推荐,随着网络信息的发展,用户以及物品的信息不断增加,协同过滤难以满足用户的个性化需求。融合 CatBoost 的协同过滤算法可以更全面地对用户及物品信息进行分析,得到更为准确的预测评分,提高用户的满意度。本算法的协同过滤算法模块,利用优化后的余弦相似度公式计算出目标用户的相似用户群,并对相似性进行排序,通过预测评分公式得到召回集并对召回集进行排序得到候选集。再利用 CatBoost 算法对数据集进行训练,对候选集进行二次评分并将两次评分结果进行融合,通过对参数以及权值的更新,达到提高算法准确度的目的,最后利用 Top-N 生成推荐列表。

2.2　算法描述

本算法包括通过协同过滤产生召回集,利用 CatBoost 进行模型训练并对召回集的物品进行二次评分,生成推荐三个阶段。算法流程如图 1 所示。

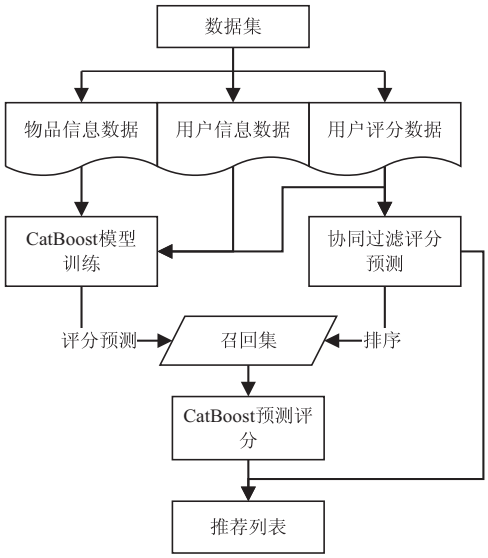

图 1　算法流程

三个阶段具体实现步骤为:

阶段 1:生成召回集 D 。

输入:用户历史评分数据。

输出:用户召回集 D 。

算法步骤:

(1)将数据集进行划分,70% 的数据作为训练集,30% 的数据作为测试集。

(2)在训练集中计算用户之间的相似度,采用加入热门商品惩罚项和时间衰减因子的优化算法复杂度的余弦相似度公式(8)进行相似度计算,并构建相似度矩阵。

(3)选取前 k 个目标用户近邻,利用式(5)计算目标用户对未评分项目的预测评分。

(4)对预测评分进行排序,组成召回集 D 。

阶段 2:CatBoost 模型训练。

输入:用户历史交互数据,用户特征数据,物品特征数据。

输出:训练好的 CatBoost 模型。

算法步骤:

(1)整合输入数据,提取特征,生成训练集。

(2)对部分特征进行优化处理,利用 CatBoost 对训练集进行模型训练。

(3)优化模型,使用 GridSearchCV 对模型进行调参。

(4)得到优化后的模型,保存模型。

阶段 3:产生推荐列表。

输入:协同过滤召回集 D ,协同过滤一次预测评分数据,CatBoost 模型。

输出:用户推荐列表。

(1)在召回集中,选取每位用户评分最高的前 k 个物品,形成候选集。

(2)利用训练好的 CatBoost 模型对候选集进行二次评分预测,与协同过滤一次评分进行加权,得到最终的预测评分,将项目最终得分进行排序。

(3)利用 Top-N 产生推荐列表反馈给目标用户。

3 实验结果与分析

3.1 实验数据集

文中采用的数据集是 MovieLens－1m 数据集，MovieLens-1M 数据集含有来自 6 040 名用户对 3 952 部电影的 100 余万条评分数据。分为三个表：用户评分信息、用户信息、电影信息。其中用户信息包括用户 id，用户性别，用户年龄，用户职业和压缩编码，其中年龄 1 对应 1 ~ 18 岁（不包含 18 岁），18 表示 18 ~ 24 岁，25 表示 25 ~ 34 岁，35 表示 35 ~ 44 岁，45 表示 45 ~ 49 岁，50 表示 50 ~ 55 岁，56 表示大于等于 56 岁，部分用户信息如表 1 所示。电影信息包括电影 id、电影名称、电影类型，部分电影信息如表 2 所示。评分数据由 6 040名用户对 3 952 部电影的评分组成，一共 1 000 209 条数据，包括用户和电影 ID，得分值以及交互信息产生的时间戳，部分评分信息如表 3 所示。

表 1 用户信息

UserID	Gender	Age	Occupation	Zip-code
1	F	1	10	48 067
2	M	56	16	70 072
…	…	…	…	…
6 040	M	25	6	11 106

表 2 电影信息

MovieID	Title	Genres
1	Toy story (1995)	Animation|Children´s|Comedy
2	Jumanji (1995)	Adventure|Children´s|Fantasy
3	Heat (1995)	Action|Crime|Thriller
…	…	…
3 952	Contender(2000)	Drama|Thriller

表 3 评分信息

UserID	MovieID	Rating	Timestamp
1	1 195	5	978 300 760
1	661	3	978 302 109
2	1 357	5	978 298 709
…	…	…	…
6 040	1 097	4	956 715 569

表 1 中的 Zip-code 为压缩编码，Occupation 为用户职业编号。

表 2 中 Genres 为电影类型特征，由符号|隔开，整个数据集共有 18 个类别特征。

Timestamp 为用户打分的时间戳，用户评分范围为 1 ~ 5，没有 0 分选择。

3.2 数据预处理

如果使用原始数据集进行集成模型训练，可利用到的特征仅有 Gender，Age，Occupation，Genres 四类。特征太少会导致模型训练出来的效果较差，所以需要从原始数据中提取出更多特征，提高数据集特征的维度。

首先处理 Genres 中的电影类型特征，将特征逐一提取出来，构造 18 * N 的特征矩阵，其中 N 为评分数据总条数，18 为特征总数。将矩阵逐列添加到评分信息文件中，列名为电影类型的名称。

从 Title 中提取出电影的相关上映时间，将上映时间加入到评分信息中，再从用户打分的时间戳中提取出用户打分的年份，将用户对电影的评分年份也作为一个特征，最后删除掉无用的信息，Zip－code，Title，Timestamp，Genres，得到最终的训练数据集。

3.3 评价标准

文中提出的算法会得到用户对于未评分物品的预测评分，所以使用均方误差（mean squared error，MSE）作为预测评分准确度指标。计算公式如下：

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{m}(y_i - y_i')^2 \tag{10}$$

均方误差可以评价数据的变化程度，MSE 的值越小，预测模型的精确度越高。

3.4 实验结果与分析

（1）Boosting 类算法准确率对比。

将数据集划分，在训练集上使用 CatBoost 算法对

数据进行训练,进行参数调节,使用 GridSearchCV 对模型进行参数调整。将 task_type 参数设置为“GPU”,可以有效提高模型的运行效率。对树的深度 depth,生成树的数量 iterations,学习率 learning_rate,子样本 subsample,对象采样方法 bootstrap_type,随机子空间方法 rsm 等参数分别进行调参设置。最终更新 depth = 10iterations = 1 100, learning_rate = 0. 15, rsm = 0. 1, subsample = 0. 66。此时得到最佳训练模型。图 3 为学习率对应的得分值,评分标准为“r2”得分值,当学习率为 0. 15 时,模型效果最好。将调节好的模型进行保存,在测试集上对模型进行效果测试。在此数据集上,分别利用三种流行的 boosting 算法进行实验,分别得到三种算法在训练集以及测试集上的结果。由图 2 的实验结果显示,在此数据集上 CatBoost 与 LightGBM 表现几乎没有差别,CatBoost 在测试集上的表现略微优于 LightGBM,提升了约 1. 5%,XGBoost 算法相对来说效果较差。以上数据在一定程度上证明了 CatBoost 算法在推荐系统中的可行性。

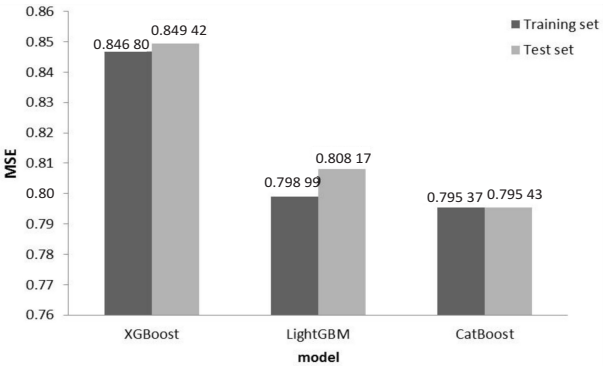

图 2　Boosting 算法准确性对比

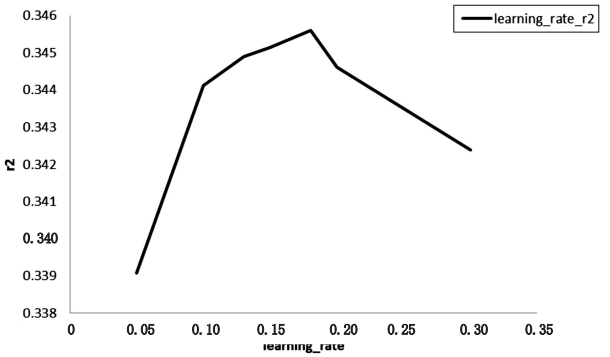

图 3　CatBoost 学习率得分

(2)召回集各算法准确率对比。

在 2.2 节算法阶段 1 中,说明了召回集的生成过程。召回集 D 共有 233 329 条数据,占总数据量的 23. 3%,混合 XGBoost 的协同过滤算法 MSE 值为 0. 800 3,文中提出的 UCF - CB 算法的 MSE 为 0. 693 15。将文中算法与文献[18]提出的 Weighted KM-Slope-Vu 算法以及文献[19]提出的 WSO 算法进行对比(见图 4),文中算法的 MSE 均小于对比算法,表明提出的推荐算法在准确性上要优于对比算法。同时相比较在测试集上的评分预测结果,UCF-XGB 与 UCF-CB 算法在准确性上都比原始的算法要有所提升。

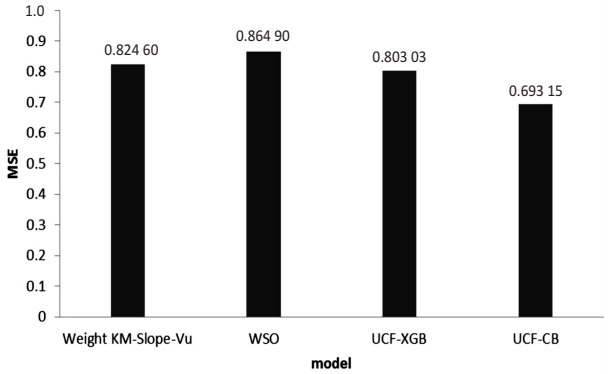

图 4　召回集上算法对比

(3)最终推荐结果分析。

为了最终产生推荐列表,本实验在召回集中对每个用户选取协同过滤评分最高的前 k 个物品,取 k 值为 8,共计 42 211 项,重新组成候选集,用混合模型进行评分预测。按照最终评分利用 Top-N 算法推荐给用户。这里的 k 值可以根据要求推荐的数量而定,要求更精确的推荐,可以适当减小 k 值,要求更广泛的推荐,可以加大 k 值。

最终的混合模型 UCF-CB 在 D_1 上的 MSE 为 0. 637 81,协同过滤算法的 MSE 为 0. 794 3,UCF-XGB 算法的 MSE 为 0. 749 67。对比原始的协同过滤算法以及 UCF-XGB 有明显的提升,文中提出的算法模型在最终的推荐集上有着较好的表现(见图 5)。

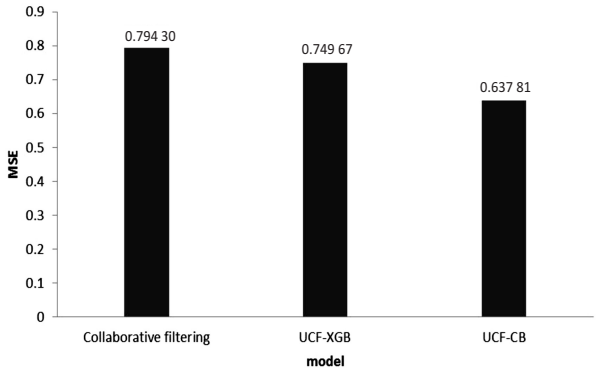

图 5　混合模型分析

4　结束语

文中提出的混合推荐算法 UCF-CB,通过改进后的协同过滤算法得到用户的召回集 D,利用训练后的 CatBoost 算法对召回集进行二次评分预测,与协同过滤一次评分进行加权融合,得到最终的预测评分。在实验 2 中,证明了该算法的优越性。最后将召回集进行压缩得到 D_1,通过 UCF-CB 算法进行评分预测,生

成推荐列表反馈给用户。文中将不同的算法进行混合,提高了推荐系统的准确性,通过实验对比验证了相比较传统的系统过滤算法有着明显的提高,并且利用 CatBoost 集成学习模型可以解决推荐系统中的冷启动问题。同时文中提出的算法也有不足之处,由于协同过滤算法在庞大的数据集上计算量过大,运行效率较差,会导致混合算法整体的效率较低。所以,下一步的工作将研究如何提高算法模型的运行效率,将矩阵分解运用到协同过滤算法中,解决数据稀疏性,对原始数据进行降维,减少计算量,使模型的运行效率更高。

参考文献:

[1] LEHMAN A, MILLER S J. A theoretical conversation about responses to information overload[J]. Information, 2020, 11(8): 379-389.

[2] BADRIYAH T, AZVY S, YUWONO W, et al. Recommendation system for property search using content based filtering method[C]//IEEE international conference on information and communications technology (ICOIACT). Bangkok, Thailand: IEEE, 2018: 25-29.

[3] XU Xiaolin, XU Guanglin. Improved collaborative filtering recommendation based on classification and user trust[J]. Journal of Electronic Science and Technology. 2016, 14(1): 25-31.

[4] KUMAR M S, PRABHU J. Hybrid model for movie recommendation system using fireflies and fuzzy c-means[J]. International Journal of Web Portals, 2019, 11(2): 1-13.

[5] ZHANG Xinran, YUAN Xin, LI Yunwei, et al. Cold-start representation learning: a recommendation approach with Bert4Movie and Movie2Vec[C]//Proceedings of the 27th ACM international conference on multimedia. Nice, France: ACM, 2019: 2612-2616.

[6] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[7] 张 锋,常会友. 使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J]. 计算机研究与发展,2006,43(4):667-672.

[8] LU C C, TSENG V S. A novel method for personalized music recommendation[J]. Expert Systems with Applications, 2009, 36(6): 10035-10044.

[9] HANCOCK J, KHOSHGOFTAAR T. CatBoost for big data: an interdisciplinary review[J]. Journal of Big Data, 2020, 7(1): 94-94.

[10] HUANG Guomin, WU Lifeng, MA Xin, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions[J]. Journal of Hydrology, 2019, 574: 1029-1041.

[11] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[M]//Theory of cryptography. New York, NY, USA: Springer, 2006: 265-284.

[12] 崔 岩,祁 伟,庞海龙,等. 融合协同过滤和 XG Boost 的推荐算法[J]. 计算机应用研究,2020,37(1):62-65.

[13] 李智彬. 基于 SVD 与 LightGBM 的音乐推荐算法研究[D]. 杭州:浙江工商大学,2018.

[14] TRUONG V H, VU Q V, THAI H T, et al. A robust method for safety evaluation of steel trusses using gradient tree boosting algorithm[J]. Advances in Engineering Software, 2020, 147: 102825.

[15] LIUDMILA P, GLEB G, ALEKSANDR V. CatBoost: unbiased boosting with categorical features[J]. Advancesin Neural Information Processing Systems, 2018(3): 6638-6648.

[16] 王 浩. 基于特征价格理论和 CatBoost 的旧机动车价值评估模型研究[D]. 天津:天津商业大学,2019.

[17] 丁 琦. 基于 Catboost 算法的员工离职预测的研究[D]. 上海:上海师范大学,2020.

[18] 袁志远. 基于协同过滤的个性化推荐算法研究[D]. 南京:南京邮电大学,2018.

[19] 温志慧. 基于项目分类和 K-means 聚类的加权 Slope One 算法研究[D]. 秦皇岛:燕山大学,2017.