计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200010

# 网上购物平台多推荐融合算法研究

朱育颉[1]　刘虎沉[2]
1 上海大学管理学院　上海 200444
2 同济大学经济与管理学院　上海 200092
(zhuyujie98@shu.edu.cn)

**摘　要**　推荐系统能帮助用户有效解决信息过载问题,现已被广泛应用于各大网上的购物平台。对用户而言,好的推荐算法能够帮助其从海量商品中快速准确发现符合自己需求的商品;对商家而言,及时呈现给用户恰当的物品能帮助商家实现精准营销,发掘长尾商品并推荐给感兴趣的用户以提高销售额。协同过滤、基于内容推荐是目前应用成熟的推荐方法,但这些方法存在数据疏散、冷启动、可扩展性差和多媒体信息特征难以提取等问题。因此,文中提出基于融合 LR-GBDT-XGBOOST 的个性化推荐算法,可有效缓解上述问题。在阿里巴巴天池大数据竞赛公开数据集上进行实验,结果显示,该算法降低了推荐稀疏性,提高了推荐精度。

**关键词**:电子商务;推荐系统;协同过滤;混合推荐
**中图法分类号**　TP18

## Research on Multi-recommendation Fusion Algorithm of Online Shopping Platform

ZHU Yu-jie[1] and LIU Hu-chen[2]
1 School of Management, Shanghai University, Shanghai 200444, China
2 School of Economics and Management, Tongji University, Shanghai 200092, China

**Abstract**　The recommender system can help users solve the problem of information overload effectively and has been widely applied in major online shopping platforms. For users, a good recommendation algorithm can help them find products which meet their needs from a large number of products. For merchants, timely presentation of appropriate items to users can help merchants achieve precision marketing, discover long-tail products and recommend them to users to increase sales. Collaborative filtering and content-based recommendation are currently mature recommendation methods, but these methods have problems such as data sparsity, cold start, poor scalability, and difficulty in extracting multimedia information features. Therefore, this paper proposes a personalized recommendation algorithm based on the fusion of LR-GBDT-XGBOOST, which can effectively alleviate the above problems. Experiments are carried out under the official dataset of the Alibaba Tianchi big data competition. The results show that the proposed algorithm reduces the recommended sparsity and improves the accuracy of the recommendation.

**Keywords**　E-commerce, Recommender systems, Collaborative filtering, Mixed recommendation

## 1　引言

近年来,随着计算机的快速普及和发展,互联网技术也得到了快速应用。同时,随着居民可支配收入的稳定增长,网络购物已经成为中国网民不可或缺的消费渠道之一。互联网提供了海量的商品选择,但这些商品的数量过于庞大,以至于消费者需要花费大量的时间和精力来甄别自己需要的商品,即产生了“信息过载”问题[1]。个性化推荐系统可以有效缓解这一情况[2]。好的推荐算法能够帮助用户快速定位目标,节约大量时间,提升用户体验。同时,它能帮助商家实现精准营销,从而提高交易量,使利润增长。

然而,传统推荐算法,例如协同过滤算法,仍存在数据疏散、冷启动[3]、可扩展性差和多媒体信息特征难以提取等问题。为了解决上述问题,学者们将用户的人口统计学信息[4]、社交信息[5]、信任度引入相似度计算中。Lei[6]分别采用 RF,LR 和 SVM 分类模型,根据用户品牌偏好进行商品筛选并预测未来一个月内用户对商品的购物行为。以上文献分别在解决数据稀疏和冷启动问题方面有所成效,但如何有效地同时解决这两个问题有待改进。为此,本文提出了一种融合 LR-GBDT-XGBOOST 的个性化推荐算法,采用真实网上购物平台数据集,对用户行为数据进行数据预处理,通过可视化分析理解内在业务逻辑,从用户、商品维度以及其中的交互关系来选择有意义的特征。实验结果显示,本文提出的融合模型的表现优于其他模型的表现。

## 2　相关算法

推荐算法的研究在国外起步较早,不少研究者从不同的角度对推荐算法进行了不同分类。从信息技术的角度分类,可依据推荐结果的算法和生成机制分为协同过滤、基于内容推荐。

基金项目:国家自然科学基金项目(61773250)
This work was supported by the National Natural Science Foundation of China(61773250).
通信作者:刘虎沉(huchenliu@tongji.edu.cn)

协同过滤算法是目前业内公认的较为成熟的推荐算法。Goldberg 等[9]首次提出协同过滤算法并将其应用于 Tapestry 电子邮件过滤系统。该算法主要根据用户之前的喜好以及与其兴趣相近的用户的选择来推荐物品。不同于 Tapestry 的单点过滤机制，Resnick 等[10]提出了跨点跨系统的新闻过滤机制 GroupLens，可自动帮助人们在大量可用文章中找到自己喜欢的文章。上述最近邻法的弊端在于需要大量计算，因此 Linden 等[11]提出基于物品的协同过滤算法。该算法通过寻找相似的商品代替寻找用户最近邻方法，在线计算花费与用户数量和物品数量无关，可以在海量数据上实时生成高质量推荐。协同过滤算法的优点较为明显，工程上实现简单、效果好、模型通用性强。但当电子商务系统规模扩大，用户、项目数据急剧增加时，数据稀疏问题便暴露。此外，它也存在较为严重的冷启动问题。Periros 等[4]将用户的人口统计信息引入到推荐算法中，从而形成混合协同过滤推荐算法，可以很好地解决冷启动问题。Shambour 等[5]在传统的基于用户的协同过滤推荐算法中融入了项目评分信任度的思想，同时摒弃了传统的相似度计算方法，实验证明该算法能缓解数据稀疏性问题。

此外，基于内容的推荐算法也能有效解决协同过滤算法整个运行过程中可能遇到的冷启动问题。在协同过滤推荐算法中，只有满足该项目被众多相关用户评价这个条件之后，才会为其他相关用户进行推荐。而在基于内容的推荐算法中，可以通过内容为用户提取特征，建立起刻画相关内容的特征向量，再根据用户的偏好去决定是否向用户推荐相关内容。Michael 等[7]将基于内容的推荐算法分成 3 个步骤：为物品提取特征表示（例如文本的 TF-IDF 向量）、特征学习、生成推荐列表。

回顾相关研究发现，针对本文所研究的网上购物平台商品推荐的具体业务场景，从算法的准确度、效率、可解释角度考虑，需要选择高效的推荐算法。此外，由于协同过滤、基于内容的推荐算法都有各自的优点及难点，实际上大多数的推荐系统都通过多种形式融合不同的推荐算法进行混合推荐，可利用各自模型推演结果进行加权组合、瀑布型混合等方式。因此，本文选取随机森林、GBDT、XGBOOST 模型作为基础模型。

### 2.1 随机森林模型

随机森林（Random Forest，RF）是由 Breiman 等[8]在分类与回归树（Classification and Regression Tree，CART）的基础上组合而成的。其原理是通过将原数据集随机采样形成不同的数据集，在单个数据集上进行决策树的训练，最终将多个决策树合并形成一个分类器，具体步骤如下。

（1）预处理 RF 模型。采用下采样方法，避免由于样本不平衡对实验结果造成影响（原数据集中正负样本比为 1∶1 100）。

1）使用 $K$ 均值聚类算法对负样本进行聚类；

2）基于相同比率的每个群集子样本，通过在随机子样本中进行测试，选择出最佳比率；

3）使用 RF 模型对下采样集进行训练和预测。

（2）参数调优。

1）对正负样本的不平衡率 $N/P$ 进行调优；

2）对森林的规模树的个数进行调优；

3）对概率阈值的设置，根据已经训练的模型输入样本特征得到预测值，预测值为概率值，默认为 0.5，通过不断调整修改概率值达到阈值的条件，从而改变对该用户和该样本的购买预测分类标签。

（3）利用 sklearn 工具包中的 RandomForest Classifier () 建立模型并进行训练，生成预测结果子集 $P$。

### 2.2 GBDT

区别于 RF 所属的 Bagging 类算法，GBDT 算法属于 Boosting 算法类。Boosting 类算法采取串形序列化模型，个体学习器之间为强关联关系，即新模型器的生成基于旧学习模型的训练结果。

GBDT 算法学习已生成 CART 分类回归树中，已有模型和实际样本输出的残差较大的样本，进而迭代生成新的 CART 树，通过上述训练方式，不断降低残差以保证结果的准确度，具体步骤如下：

（1）使用 $K$ 均值聚类对负样本进行聚类。

（2）基于相同比率的每个群集子样本，通过在随机子样本中加入 GBDT 进行测试，选择出最佳比率。

（3）为 GBDT 分类器选择最佳参数。

1）对正负样本的不平衡率 $N/P$ 进行调优；

2）为 GBDT 选择最佳的森林规模树的个数和学习率；

3）为基础树选择最佳的最大深度值、内部节点再划分所需最小样本数和叶子节点最少样本数；

4）调优概率阈值。

（4）利用 sklearn 工具包中的 GradientBoostingClassifier () 建立模型并进行训练，生成预测结果子集 $P$。

### 2.3 XGBOOST

同属于 Boosting 类算法，XGBOOST 高效实现并优化了 GBDT 算法。在基学习器的选择上，XGBOOST 模型既可以采用树模型，也可以采用其他模型（例如 LR），因此不局限于 GBDT 算法限定的 CART 树。XGBOOST 通过树的不断特征裂变来学习新函数和拟合残差，具体步骤如下。

（1）训练 XGBOOST 模型参数进行分析：

1）用较高的学习率调整最佳的森林规模树的个数；

2）调整最大深度值和子节点中样本权重和的最小值；

3）调整参数 $\gamma$；

4）调整 $\lambda$ 和 $\alpha$；

5）降低学习率并循环以获得更稳定的参数组合。

（2）利用 sklearn 工具包中的 GradientBoostingClassifier () 建立模型并进行训练，生成预测结果子集 $P$。

## 3 数据处理

### 3.1 数据源介绍

本文数据源于阿里旗下天池大数据平台竞赛，通过网上购物平台真实脱敏数据构建商品推荐模型。数据集包含用户在 30 天时间内的移动端行为数据。用户表中包含 6 个字段，分别为用户标识、品牌标识、用户位置标识、商品分类标识、用户对商品的交互行为和行为时间。具体数据量如表 1 所列。

表 1 数据集描述

Table 1 Dataset description

| 数据集 | 数量 |
|---|---|
| 用户数 | 91027 |
| 商品子集 | 62098 |
| 用户商品交互数据 | 23201027 |

### 3.2 数据预处理

（1）剔除异常时段的数据

由于本文数据中包含淘宝 12 月 12 日购物节的数据，对

这一个月来用户的总操作(浏览、收藏、加入购物车、购买行为)进行分析发现,购买行为在12日0点翻倍式增长,明显为异常数据,因此将当日数据剔除。

(2)剔除重复值、无效值

在电商网站数据统计过程中,可能由于统计存在重复数据、异常数据等问题数据,因此本阶段的数据清洗环节将对上述重复值、无效值进行直接剔除。

(3)平衡样本数据

由于原始数据集正负样本比值约为1∶1100,极为失衡,因此可能会导致算法预测模型将正样本集看作噪音数据,从而偏向负样本集数据。相比负样本集,正样本集被错分的几率更大。因此,本文通过随机下采样方法来避免以上问题。

### 3.3 特征构建

根据前文数据预处理结果并结合网上购物平台业务特点重构数据特征。本文主要从用户特征、商品特征、商品类目特征这三大维度及其组合完成特征构建。将本实验视作一个二分类问题,则输出变量 $Y$ 标签分别为1(购买)和0(未购买)。

(1)用户行为特征

1)用户活跃度。其指用户在最近 $N$ 天的行为总和,反映了用户的购买习惯。

2)用户转换率。其反映了用户的购买决策操作习惯。

(2)商品特征

1)商品活跃度。其指前 $N$ 日用户对类别操作的行为总和,反映了该商品的热度。

2)商品的转换率。其反映了商品的购买决策操作特点(例如价值高、决策时间长)。

(3)商品类目特征

1)商品类目活跃度。其指前 $N$ 日用户对该类目操作行为总和,反映了该类目的热度。

2)商品类目购买转换率。其反映了商品类目的购买决策操作特点。

(4)用户-商品类目特征

1)用户-商品类目浏览总和;

2)用户-商品类目购买总和;

3)用户-商品类目最后操作日期;

(5)商品-商品类目特征

1)商品在所属类目中用户行为总和排序;

2)商品在所属类目中的销量排序。

### 3.4 划分数据集

本实验将数据集划分为两部分,包括训练集和测试集。学习训练过程如图1所示。

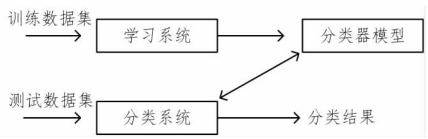


图1 数据模型学习训练的过程

Fig. 1 Process of training data model

数据预处理结果显示,本数据集的用户点击转化率不足1%,即平均每个用户的100次交互行为(包括浏览、点击、加购商品)才能促成1次交易行为,数据量较为庞大。若将用户全时间段所有交互行为都作为特征数据,一是会极大地降低模型的计算速度,二是由于用户兴趣具有时间衰退性,将会降低模型的准确度。因此,在划分数据集时选择以周为间隔单位,可分为4组,其中涉及到一组异常数据,在3.2节已说明剔除原因。

## 4 算法性能分析

### 4.1 算法评估指标

当用户评分稀疏或缺失时,可考虑使用F1-score这一综合指标,其特点在于调和精确度(Precision)和召回率(Recall),三者的基础含义为:

$$Precision=\frac{TP}{TP+FP} \tag{1}$$

$$Recall=\frac{TP}{TP+FN} \tag{2}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{3}$$

其中涉及到混淆矩阵的相关元素,如表2所列。

表2 混淆矩阵

Table 2 Confusion matrix

| | 预测正类 | 预测负类 |
|---|---|---|
| 正类样本 | TP(True Positive) | FN(False Negative) |
| 负类样本 | FP(False Positive) | TN(True Positive) |

结合网上购物平台的商品推荐场景,精确度(Precision)和召回率(Recall)又可表示为:

$$Precision=\frac{推荐商品数量\cap 有用商品数量}{推荐商品数量} \tag{4}$$

$$Recall=\frac{推荐商品数量\cap 有用商品数量}{有用商品数量} \tag{5}$$

精确度高则推荐更精准,召回率高则推荐的接受度更高。只有当两者都较高时,才能说明推荐算法的性能较好,因此,将F1-score作为主要指标。

### 4.2 算法实验结果分析

#### 4.2.1 实验环境

本实验的操作系统为Windows 7-32位,处理器为Intel(R) Core(TM) 15-5200U CPU @ 2.20 GHz (4 CPUs),2.2 GHz。使用的python版本为3.7.2,并基于pandas、sklearn等工具包实现。

#### 4.2.2 独立模型训练结果

首先训练LR模型,但结果不理想。在采用RF,GBDT及XGBOOST时效果提升明显,这3种集成算法中,在XGBOOST与RF的对比中,虽然XGBOOST只在F1评分上有轻微优势,但在模型实现用时方面比RF更短。LR模型简单,运行时间短,但准确性对比其余算法表现较差。本文中这3种树集成模型(RF、GBDT、XGBOOST)对此类特征较为复杂的数据,显现出了较强的泛化能力。

表3 独立模型训练结果

Table 3 Independent model training result

| 模型 | F1评分 | 准确率 |
|---|---|---|
| LR | 0.0509849 | 0.06358382 |
| RF | 0.0887884 | 0.09532374 |
| GBDT | 0.0848387 | 0.09947368 |
| XGBOOST | 0.0893785 | 0.08753677 |

4.2.3 混合模型训练结果

由于各种推荐算法都有各自的优点及实现难点，实际上大多数的推荐系统通过多种形式融合不同的推荐算法进行混合推荐：

(1)加权混合。加权混合组合是一种较为常见的方式。可利用各自模型的推演结果进行加权组合将推荐结果及分数组合生成最终推荐集。通过合理的加权方式，综合多种算法的结果较优，但也存在着运算量大、系统复杂等问题。

(2)瀑布型混合。瀑布型混合方式采用“过滤”原理，将一个模型的过滤结果作为另一个模型的输入数据。通常以 GBDT+LR 模型为例，结合 LR 模型与 GBDT 模型各自的优点，将 GBDT 算法输出结果直接输出为 LR 输入特征，层层递进，提高数据信息利用率且提升 LR 模型的学习效率。

(3)分级混合。工业中，通过此种混合方式，使混合算法变得简单高效。结合不同场景，根据不同算法的特点首先使用高精度算法，用其他算法得出后续结果。

表 4 混合模型的训练结果

Table 4 Mixed model training result

| 模型 | F1 评分 | 准确率 |
|---|---|---|
| LR+GBDT | 0.1015031 | 0.06358382 |
| RF+XGBOOST | 0.0987884 | 0.09532374 |
| GBDT+XGBOOST | 0.0893785 | 0.08578942 |
| LR+GBDT+XGBOOST | 0.1105109 | 0.08051091 |

在实验中，根据上一步模型的独立训练结果进行模型间的融合。实验中涉及到大量数据处理、特征选取、参数调节的步骤，通过实验得到如下结论：

(1)正常情况下，模型间的融合会提升性能。但当单模型表现较差时，与其他模型融合反而会降低与其融合模型的效果。

(2)在混合推荐中，直接融合效果较差，例如简单的加权可采用分级策略，将一个模型的结果输出作为另一个模型的输入，从而提升效率。

(3)模型间的实验性能相似时，融合时的效果也不会有显著提升。

(4)最终结果显示，LR，GBDT 和 XGBOOST 相融合的模型的效果最佳。

**结束语** 本文对网上购物平台推荐算法进行深入探究，设计并实现了一个融合优化多种推荐算法的模型，在数据集上的实验结果表现良好。本文的后期研究可以从实时推荐、可视化推荐结果生成、结合场景挖掘用户兴趣模型等方向入手。

## 参 考 文 献


[1] BORCHERS A，HERLOCKER J. Ganging up on information overload[J]. Computer，1998，31(4)：106-108.

[2] ADOMAVICIUS G，TUZHILIN A. Toward the next generation of recommender systems：a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering，2005，17(6)：734-749.

[3] YU H，LI J H. A recommendation algorithm to solve the cold start problem of new projects[J]. Journal of Software，2015，26(6)：1395-1408.

[4] PEREIRA A L V，HRUSCHKAE R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems[J]. Knowledge-Based Systems，2015，82：11-19.

[5] SHAMBOUR Q，LU J. An effective recommender system by unifying user and item trust information for B2B applications[J]. Journal of Computer and System Sciences，2015，81(7)：1110-1126.

[6] LEI M L. Research on shopping behavior based on Alibaba Big Data[J]. Internet of Things Technology，2016，6(5)：57-60.

[7] PAZZANI M J，BILLSUS D. Content-Based Recommendation Systems[C]// Adaptive Web. Springer-Verlag，2007：325-341.

[8] BREIMAN L，BREIMAN L，CUTLERR A. Random Forests Machine Learning[J]. Journal of Clinical Microbiology，2001，2：199-228.

[9] GOLDBERG D，NICHOLS D A，OKI B M，et al. Using collaborative filtering to weave an information TAPESTRY[J]. Communications of the ACM，1992，35(12)：61-70.

[10] RESNICK P，IACOVOU N，SUCHAK M，et al. GroupLens：an open architecture for collaborative filtering of netnews[C]// Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. 1994：175-186.

[11] VERBERT，MANOUSELIS N，OCHOA X，et al. Context-Aware Recommender Systems for Learning：A Survey and Future Challenges[J]. IEEE Transactions on Learning Technologies，2012(4)：318-335.


**ZHU Yu-jie**，born in 1998，postgraduate. Her main research interests include machine learning and so on.