DOI：10.16667/j.issn.2095-1302.2017.02.030

XGBoost算法在电子商务商品推荐中的应用

张 昊[1]，纪宏超[2]，张红宇[1]
（1. 中南大学 商学院，湖南 长沙 410083；2. 中南大学 化学化工学院，湖南 长沙 410083）

摘 要：近年来，在电子商务网站进行在线购物已逐渐成为人们主要的购物途径之一。在在线购物过程中，人们会留下大量的浏览信息，但只有极少数会转化为购买。对用户信息进行数据挖掘，个性化向用户推荐商品可以有效提高用户的购物效率并提高商家的收入。XGBoost算法是一种高效准确的分类算法，文中将XGBoost算法应用于商品推荐中，从而实现了准确预测用户购买行为的目的，为商品推荐提供了一种有效的方法。

关键词：电子商务；大数据；推荐算法；分类
中图分类号：TP39　　**文献标识码：**A　　**文章编号：**2095-1302（2017）02-0102-03

0 引 言

近年来，在电子商务网站进行在线购物逐渐成为人们新的购物习惯。在线购物过程中，人们在最终决定购买某种商品前，通常会在电子商务平台留下大量的信息，这些信息通常反映了用户购物的行为模式。通过数据挖掘方法来分析用户的行为模式数据，有利于更好地了解用户的购物习惯和倾向性，从而为预测用户的购买行为提供可能[1]。准确预测用户的购物行为对电子商务平台而言具有重要意义，通过预测结果可以个性化地向用户推荐商品，提高用户的购物效率，促成更多交易，提高营业收入。因此，国内外大型电子商务企业都不同程度上运用了商品推荐算法，学者也将统计和机器学习方法用于商品推荐的研究中，以期提高预测的准确度。

雷名龙[2]分别采用随机森林、逻辑回归和SVM分类模型，以阿里巴巴电子商务平台4个月的购物数据为研究对象，对用户未来是否会购买某种商品做出行为预测，超过5%的准确率。张春生等[3]考察了品牌可信度、价格、付款人数等多种评价指标对于用户购买行为的相关性。Vieira等[4]采用深度置信网络和自编码器等深度学习策略，就筛选出的商品及用户特征进行建模，将其与传统的决定树、随机森林等算法进行比较，发现深度学习方法有利于获得更好的预测结果。马月坤等[5]采用构建用户行为知识库的方法，对客户的行为信息进行了有效存储和更新管理。

eXtreme Gradient Boosting（XGBoost）[6]是一种基于梯度Boosting的集成学习算法，其原理是通过弱分类器的迭代计算实现准确的分类效果。梯度Boosting因其分类的高效性和准确性被广泛应用于人脸识别[7]、火灾识别[8]、列车停车[9]等诸多方面。本文将XGBoost引入到电子商务的商品推荐算法中，挖掘用户在电子商务平台的行为数据信息，建立分类预测模型，从而个性化地向用户推荐商品。结果表明，与传统机器学习算法相比，XGBoost具有速度快、准确度高等优势。

1 数据描述

本文所使用的数据来自阿里巴巴天池大数据竞赛公开数据集，包含20 000名用户某年11月18日至12月18日一个月的完整行为数据。每条购物行为包含4个字段，分别为用户ID、品牌ID、用户对商品的交互行为和行为时间。用户与商品的交互行为分为“浏览”、“收藏”、“加入购物车”和“购买”。

1.1 异常值剔除

异常值的存在通常会严重影响建模和预测质量[10]，因此有必要对数据中存在的异常值进行剔除。获取的数据时间内包含淘宝“双12”购物节，当日用户的总浏览、收藏、加入购物车和购买总次数分别为往日均值的1.8、1.4、2.4和4.5倍，属于明显异常值，因此当日的全部数据在后续处理过程中被剔除。此外，在1个月内无购买记录的用户可能不具备在线购物习惯，此类用户对于预测建模不具备参考价值，因此此类数据也被剔除。

1.2 特征筛选

原始数据无法直接用于建模，因此需要将其归纳为统计特征。特征的筛选需要能够充分描述商品信息、用户信息及用户-商品的交互情况。因此我们使用的特征如表1所列。

在表1中，商品特征主要反映了商品的热度，通常交互和购买次数高的商品具有更高的性价比，因此能够吸引用户的购买。用户特征则主要反映了用户的购物习惯，如其购物频率以及用户更多选择冲动购物还是反复迟疑后才会购买。交互特征则更多考虑到用户与商品之间的交互行为。通常在购物过程中，用户会将某商品与同类商品比较后才会选择是否购买，因此用

收稿日期：2016-10-08
基金项目：中南大学国家级大学生创新创业资助项目

户与同类商品的交互行为也应被考察。

表 1 用于建模的统计

特征描述	特征种类	编号
用户1日内对该商品的各类交互次数	交互特征	1 - 4
用户1日内对同类商品的各类交互次数	交互特征	5 - 8
用户3日内对该商品的各类交互次数	交互特征	9 - 12
用户3日内对同类商品的各类交互次数	交互特征	13 - 16
商品总浏览（收藏、加入购物车、购买）量	商品特征	17 - 20
商品最近3日内浏览（收藏、加入购物车、购买）量	商品特征	21 - 24
用户总浏览（收藏、加入购物车、购买）量	用户特征	25 - 28
用户浏览（收藏、加入购物车）购买量比	用户特征	29 - 31

1.3 样本划分

由于数据总量较大，在处理过程中仅使用部分样本进行建模。同时，第 25 天的数据由于异常值已被剔除，应尽量消除其影响。因此，我们选择第 8、15、22 天的数据，每天抽取 2 万个样本作为训练集。训练集的每个样本由一个用户 - 商品对组成。特征的统计涉及到前三天的信息，因此这样划分样本比较具有代表性。选择第 23 天的 6 万个样本作为测试集。在训练集的 6 万个样本中，阳性样本为 100 个，而在测试集的 6 万个样本中，阳性样本为 112 个。可以看出，样本具有高度的不平衡性。这是因为用户会浏览大量的商品，但其中转化为实际购买的仅为其中极少的一部分。

2 分类建模

2.1 XGBoost 算法

Boosting 是一种非常有效的集成学习算法，采用 Boosting 方法可以将弱分类器转化为强分类器，从而达到准确的分类效果。其步骤如下所示：

（1）将所有训练集样本赋予相同权重；

（2）进行第 m 次迭代，每次迭代采用分类算法进行分类，采用公式（1）计算分类的错误率：

$$\mathrm{err}_m = \frac{\sum \omega_i I(y_i \neq G_m x_i)}{\sum \omega_i} \tag{1}$$

式中 ω_i 代表第 i 个样本的权重，G_m 代表第 m 个分类器；

（3）计算 α_m=log（（1 − err_m）/err_m）；

（4）对于第 m+1 次迭代，将第 i 个样本的权重 ω_i 重置为 $\omega_i \times e^{\alpha_m \times I(y_i \neq G_m x_i)}$；

（5）完成迭代后得到全部的分类器，采用投票方式得到每个样本的分类结果。其核心在于每次迭代后，分类错误的样本都会被赋予更高的权重，从而改善下一次分类的效果。

Gradient Boosting 是 Boosting 的一个改进版本，经证明，Boosting 的损失函数是指数形式 [11]，而 Gradient Boosting 则是令算法的损失函数在迭代过程中沿其梯度方向下降，从而提升稳健性。其算法流程如下所示：

（1）初始化 $f_0(x) = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, \gamma)$

（2）对于 1 − m 次迭代：

① $F_0(x) = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$

② 对于 m=1 到 M：

$$\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \ldots, N$$

$$a_m = \mathrm{argmin}_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(X_i; a)]^2$$

$$\rho_m = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(X_i) + \rho h(X_i; a_m))$$

$$F_m(X) = F_{m-1}(X) + \rho_m h(X; a_m)$$

XGBoost[12] 是一种 Gradient Boosting 算法的快速实现，它能够充分利用多核 CPU 进行并行计算，同时在算法上进行改进以提高精度。本文采用 XGBoosting 算法的 R 语言版本进行分类建模，采用 10 折交互检验优化参数。

2.2 其他分类方法

为便于比较，我们也采用了另外两种通用的分类方法——分类树和随机森林作为比较。分类树（CART）最早由美国斯坦福大学和加州大学伯克利分校的 Breiman 等人于 1984 年提出。分类树采用二元递归划分方法，构建二叉树从而处理二分类问题，具有原理简单、速度快等优点。但其分类准确程度较差，容易出现过拟合。随机森林的原理是随机建立大量的分类树，每棵树单独对样本进行分类，最终分类结果由每棵树各自的分类结果通过投票确定。随机森林算法提高了分类的准确性，且结果稳健，易于调整参数，但运行速度较慢。

2.3 评价标准

对于商品推荐算法，我们更关注的是对于阳性样本的预测是否准确，因此采用阳性样本预测的准确率（precision）、召回率（recall）和 F_1 值作为评价指标，其定义如下：

$$\mathrm{precision} = \frac{T_P}{T_P + N_P} \tag{2}$$

$$\mathrm{recall} = \frac{T_P}{P} \tag{3}$$

$$F_1 = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{4}$$

其中，P 为阳性样本总数，T_P 为正确预测的阳性样本数量，N_P 为错误预测的阳性样本数量。同时，对于大量样本的数据处理，运算速度也是重要的评价指标。本实验在个人计算机（CPU：Intel i7 4710MQ 2.3 GHz；RAM：16 G）上运行，用 R 语言的 system.time（）函数记录运行时间。

3 结果与讨论

3.1 结果比较

通过交互检验分别优化三种算法的参数后，采用测试集样本对建模结果进行预测，三种算法的结果如表 2 所列。

结果表明，在三种分类算法中，XGBoost 的预测结果较随机森林略高，但运行速度要显著快于随机森林，而分类树原理简单，运算速度与 XGBoost 接近，但运算准确性明显较差。因此 XGBoost 相较于另外两种算法具有准确性高、运算速度

快等优势。

表 2 分类结果比较

模型名称	准确率	召回率	F_1值	运算时间/s
XGBoost	0.99	0.88	0.93	1.12
随机森林	1.00	0.79	0.88	6.16
分类树	0.90	0.16	0.27	1.68

3.2 变量重要性分析

通过 XGBoost 和随机森林的建模可以判断每个特征变量对模型的贡献程度，从而判断哪些特征变量对于用户购买行为的影响更为显著。分析结果如图 1 所示。

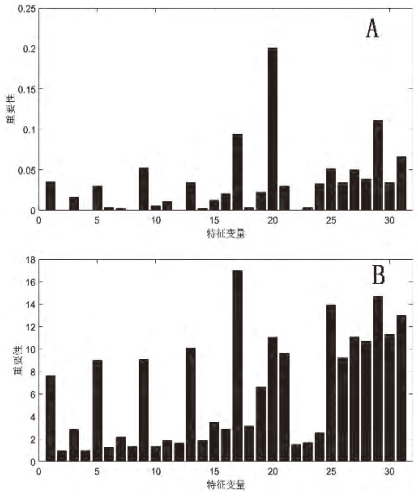

图 1 特征变量重要性

图 1(A) 所示为 XGBoost 模型的变量重要性结果，图 1(B) 所示为随机森林的变量重要性结果，其中，第 17、21、25 和 29 个变量在两个模型中的重要性排序中均在前四位，其对应的特征分别为商品的总浏览量、商品最近三日浏览量、用户总浏览量和用户浏览 / 购买比。与浏览动作相关的变量对模型的贡献程度最大，是因为浏览是用户与商品交互的最主要方式，其信息丰富程度远高于其它特征。除浏览外，与用户“加入购物车”这一动作相关的特征最高，因为用户将商品加入购物车后，很可能在未来几日内进行购买。

在用户 - 商品交互特征中，我们发现用户三日内与商品交互相关的特征变量，重要程度并不低于用户前 1 天的商品交互，这说明用户购买商品前有一定的犹豫时间。用户与同类商品交互相关的特征变量的重要程度略高于用户与某一商品的交互，这说明大多数用户在购买某商品时，会将其与同类商品进行充分比较，并最终选择那些被浏览和购买次数较多的热门商品。

此外，用户特征在重要性排序中的位置均较高，意味着不同的用户有着不同的购物习惯，因此，针对不同用户进行更加个性化的推荐非常必要。

4 结 语

本文采用 XGBoost 分类算法，基于阿里巴巴的真实用户数据进行了特征提取和分类建模，并与随机森林、决策树算法进行了对比，得到了较准确的预测结果。通过对变量重要性进行分析，我们识别了对模型贡献较高的变量。该研究有助于理解用户浏览信息与其购买行为的相关性，对个性化推荐系统性能的完善有重要的现实意义。

参 考 文 献

[1] Chester Curme, Tobias Preis, H. Eugene Stanley, et al. Quantifying the semantics of search behavior before stock market moves[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111 (32) : 11600-11605.
[2] 雷名龙 . 基于阿里巴巴大数据的购物行为研究 [J]. 物联网技术，2016，6 (5) : 57-60.
[3] 张春生，图雅，翁慧，等 . 基于电子商务同类商品的推荐算法研究 [J]. 计算机技术与发展，2016，26 (5) : 17-21.
[4] Vieira A. Predicting online user behaviour using deep learning algorithms[J].Computer Science, ArXiv151106247 Cs Stat, 2015, 46 (8) : 127-135.
[5] 马月坤，刘鹏飞 . 基于知识库的客户网购意向预测系统 [J]. 计算机工程与应用，2016, 52 (13) : 101-109.
[6] Friedman J H. Greedy Function Approximation : A Gradient Boosting Machine[J]. Annals of Statistics, 2000, 29 (5) : 1189-1232.
[7] 杜晓旭 . 基于 Boosting 算法的人脸识别方法研究 [D]. 杭州 : 浙江大学，2006.
[8] 杨国田，吴章宪，杨鹏远 .Boosting 在火灾识别中的应用研究 [J]. 计算机工程与应用，2010，46 (5) : 200-204.
[9] 周骥，陈德旺 . 机器学习在列车精确停车问题的应用 [J]. 计算机工程与应用，2010，46 (25) : 226-230.
[10] 张德然 . 统计数据中异常值的检验方法 [J]. 统计研究,2003,20(5): 53-55.
[11] Theofanis Sapatinas. The Elements of Statistical Learning[M]. Springer, 2001 : 192-192.
[12] Chen T, He T, Benesty M. xgboost : Extreme Gradient Boosting[J].2016, 5 (9) : 222-208.
[13] Mathias M. Adankon, Mohamed Cheriet. Support Vector Machine[J].Computer Science, 2002, 1 (4) : 1-28.
[14] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3 (4) : 212-223.

（上接第 101 页）

民邮电出版社，2008.
[5] 陈刚 . 基于 SSH 的 J2EE 开发平台研究与应用 [D]. 成都 : 四川师范大学，2007.
[6] 孟强，单玉祥，李阳冬，等 . 基于短距离无线通信的交通信息检测系统设计与实现 [J]. 物联网技术，2015，5 (7) : 14-15.
[7] 田晓娜，赵晴 . 基于 SSI 框架的考勤系统的设计与实现 [J]. 物联网技术，2015，5 (2) : 76-77.
[8] 唐永瑞，张达敏 . 基于 SSI 的应急事务管理系统的设计与实现 [J]. 计算机技术与发展，2014 (4) : 151-154.

作者简介：谭善伟（1990—），男，湖南人，贵州大学控制工程专业2014级研究生。研究方向为控制理论及应用。