Advances in Applied Mathematics 应用数学进展, 2022, 11(3), 923-934
Published Online March 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.113099

Hans汉斯

基于BalanceCascade的软投票策略的商品销售预测方法

张　晨，杨　进

上海理工大学，理学院，上海

收稿日期：2022年2月6日；录用日期：2022年3月1日；发布日期：2022年3月8日

摘　要

随着互联网发展，网上购物已经成为人们生活中不可或缺的一部分，为了实现更好的帮助顾客推荐商品的目的。首先根据原有数据生成新的特征值，再用互信息的方法对数据进行特征选择；其次利用BalanceCascade算法处理类别不平衡的问题，借助集成策略弥补欠采样的缺陷，与简单采样方法相比，能够对样本数据得到充分的利用还降低了正负样本差造成的影响；最后选择用软投票的方法将XGBoost和随机森林结合为一个分类器做预测，降低了单一的算法所造成的偏差，从而得到更好的结果。基于阿里巴巴天池大赛所提供的数据，以查准率P、召回率R和F1值为评价指标，分别与当前热门的机器学习算法和融合模型进行对比，验证了该方法的有效性。

关键词

互信息，类别不平衡，软投票，随机森林，极限梯度提升

Commodity Sales Forecast Method Based on BalanceCascade Soft Voting Strategy

Chen Zhang, Jin Yang

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 6th, 2022; accepted: Mar. 1st, 2022; published: Mar. 8th, 2022

Abstract

With the development of the Internet, online shopping has become an indispensable part of people's life, in order to achieve a better purpose to help customers recommend products. Firstly, new eigenvalues are generated according to the original data, and then the data are selected by

文章引用: 张晨, 杨进. 基于 BalanceCascade 的软投票策略的商品销售预测方法[J]. 应用数学进展, 2022, 11(3): 923-934. DOI: 10.12677/aam.2022.113099

the method of mutual information. Secondly, the BalanceCascade algorithm is used to deal with the class imbalance, and the integration strategy is used to make up for the defect of undersampling. Compared with the simple sampling method, it can make full use of the sample data and reduce the influence of positive and negative sample difference. Finally, the Softvoting method is used to combine XGBoost and random forest into a classifier to make prediction, which reduces the deviation caused by a single algorithm and gets better results. Based on the data provided by Alibaba Tianchi Competition, the accuracy rate P, recall rate R and F1 values are compared with the current popular machine learning algorithms to verify the effectiveness of this method.

Keywords

Mutual Information, Class-Imbalance, Softvoting, Random Forest, Extreme Gradient Boosting

1. 引言

随着我国国民经济的飞速发展，人民的生活水平明显提高，以电子商务为代表的新兴产业快速发展。购物网站聚集了海量的商品和用户信息，用户可以通过网站的商品搜索，足不出户，及时方便的获取海量的商品和用户信息，购买心仪的商品，而不是局限于线下购买这种传统的交易方式，极大的方便了用户的购买体验。本文主要解决的是消费者网上购买行为预测问题。希望网站能够根据用户的喜好推荐用户感兴趣的商品，提高用户的体验，改善互联网的购物环境。

对于该问题国内方面，李旭阳[1]等人提出了长短期记忆网络(Long Short-Term Memory, LSTM)和随机森林(Random Forest, RF)相结合的预测模型进行研究，它们从原始数据中分别提取静态特征和动态特征，对这些特征通过 LSTM 进行特征选择，再用随机森林算法做预测。马倩[2]先通过特征工程从原始数据中提取出客户和商品的相关特征，接着对基于决策树的极限梯度提升算法(Extreme Gradient Boosting, XGBoost)和逻辑回归算法(Logistic Regression, LR)分别用人工赋权重法和线性模型学习赋权重法融合后对商品平台商品复购做预测。陈龙[3]通过当前较为流行的三种梯度提升算法分别用投票法和堆叠法做融合后对商品购买行为做预测。张震[4]通过对装袋算法(Bootstrap aggregating, Bagging)方法改进提出了一种细分化集成模型融合单一算法，提高了商品复购行为的预测准确率。邹润[5]对商品平台关于商品的点击数据设计了直接特征、转化率特征、时间特征和转化率特征，用随机森林的方法做特征选择后，再用渐进梯度回归树的方法取得较好的预测结果。

国外方面，Y Tian 等人根据个人与个人之间的电子商务(Consumer to Consumer, C2C)线上线下购销量具有显著差异的缺点，构建了一种使用用户交易频率和时间的预测模型，通过在真实交易数据中验证其准确性。通过传统的 Pareto/NBD 模型，为预测 C2C 电子商务下用户重复购买行为提供了一个简单而强大的工具[6]。Zuo Y 等人通过采用支持向量机(support vector machines, SVM)模型处理消费者在超市产生的射频识别数据，研究发现 SVM 模型与线性回归等预测模型相比，显著提高了消费者购买行为的预测准确率[7]。HJ Chang 等人通过聚类分析和关联规则分析提出了潜在客户购买行为的预测模型，其中聚类分析提取搜集忠诚客户的个人信息数据用于定位潜在客户属性，关联规则分析提取忠诚客户购买行为的特征用于检测客户对热销商品的兴趣[8]。YS Cho 等人提出了一种基于最近一次消费、消费频率和消费金额，

即 RFM (Recency, Frequency, Monetary)模型的新增量加权挖掘方法，用于消费者购买行为中的预测，并为了验证其有效性，收集了网上商城的化妆品用户消费行为数据集中进行实验[9]。

本文通过阿里巴巴天池大赛所提供的数据进行数据处理后做特征工程，生成新的衍生特征，再通过互信息的方法进行特征选择，通过文章提出的 BalanceCascade [10]采样方法对数据进行采样，用软投票策略将分类算法融合后进行预测，将预测结果与传统的机器学习方法做出的预测结果作对比，也将简单欠采样，EasyEnsemble [11]和 BalanceCascade 的采样方法做出的预测结果作对比，得出结论。

2. 特征模型

本文所使用的数据来源于阿里巴巴提供的 2014 年 11 月 18 日到 2014 年 12 月 18 日的 10,000 名用户在某网站上的用户与商品的历史交互记录。但是无法从这些数据中得到实用的特征数据，所以我们必须对原始数据进行特征提取。原始数据有如下特征。

Table 1. Original features
表 1. 原始特征

特征类别	特征名称	特征描述
用户特征	User_id	用户标识
	User_geohash	用户地址
商品特征	Item_id	商品标识
	Item_category	商品类别
用户－商品特征	Behavior_type	用户对商品的交互行为
	Time	用户对商品的交互行为的发生时间

文章根据表 1 所给出的原始特征来提取新的特征，同样也把它们分为用户特征，商品特征和用户－商品特征 3 个特征类别。

2.1. 特征提取算法

算法 1：特征提取算法

输入：天池大赛记录的数据

输出：商品特征，用户特征，商品－用户特征

(a) 统计测试样本和训练样本的数据中发生购买行为和未发生购买行为的数据 item_id 和 user_id；

(b) 遍历样本数据之前的 1 天，3 天，5 天，7 天，9 天内每个 item_id 发生的购买，浏览，收藏，加购物车的行为次数；

(c) 遍历样本数据之前的 1 天，3 天，5 天，7 天，9 天内每个 user_id 发生的购买，浏览，收藏，加购物车的行为次数；

(d) 遍历样本数据之前的 1 天，3 天，5 天，7 天，9 天内每个 user_id 对应的 item_id 发生的购买，浏览，收藏，加购物车的行为次数；

(e) 分别构造商品特征，用户特征，商品-用户特征，即将(b)，(c)和(d)所统计的行为次数进行加权或作相除形成新的特征。

2.2. 用户特征

用户特征保留了原始特征 User_id，而对于特征 User_geohash 缺失数据超过了总数据的三分之二，并

且对数据进行了加密处理，因此数据间不具有相关性，所以本文决定把该特征值剔除掉。新加入的用户特征如表 2。

Table 2. User features
表 2. 用户特征

特征类别	特征名称	特征描述
用户特征	user_active	用户的活跃度
	userbuy_browse_ratio	用户购买浏览比值
	userbuy_addcart_ratio	用户购买加购车比值
	userbuy_favior_ratio	用户购买收藏比值
	userbrowse_num	用户浏览所有商品的数量
	userfavior_num	用户收藏所有商品的数量
	useraddcart_num	用户加购所有商品购物车的数量
	userbuy_num	用户购买所有商品的数量

2.3. 商品特征

商品特征保留了原始特征 item_id 和 item_category，增添了新的商品特征，新加入的特征如表 3。

Table 3. Item features
表 3. 商品特征

特征类别	特征名称	特征描述
商品特征	item_active	商品热度
	itemcategory_active	商品种类热度
	itembrowse_num	商品被浏览的数量
	itemfavior_num	商品被收藏的数量
	itemaddcart_num	商品被加购物车的数量
	itembuy_num	商品被购买的数量

2.4. 用户－商品特征

对于用户－商品特征，本文将用户购买商品的时间划分为日期和具体的时刻，用户－商品的原始特征无法作为所需要的特征直接运用，但是用户特征，商品特征以及用户－商品特征当中的新特征都是从特征 Behavior_type 和 Time 中提取出来的。新加入的用户－商品特征如表 4。

Table 4. User-Item features
表 4. 用户－商品特征

特征类别	特征名称	特征描述
用户－商品特征	item_user_active	用户对商品的喜爱程度
	item_user_shop	用户对商品的购买行为
	item_userbrowse_num	某用户对某商品浏览数量
	item_userfavior_num	某用户对某商品收藏数量
	item_useraddcart_num	某用户对某商品加购物车数量
	item_userbuy_num	某用户对某商品购买数量

由于不同的时间刻度对当前的预测影响不同，所以文章取了 5 个时间刻度 9 天，7 天，5 天，3 天，1 天，对不同的时间刻度又形成了新的不同特征，再加上保留的部分特征，总共有 108 维特征。

3. 算法介绍

3.1. 互信息算法

互信息[12]是信息论里一种有用的信息度量，它可以看成是一个随机变量包含另一个随机变量的信息量。

设两个随机变量(X,Y)的联合分布为$p(x,y)$，边缘分布分别为$p(x)$，$p(y)$，互信息$I(X;Y)$是联合分布$p(x)$ $p(y)$的相对熵，即

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x),p(y)} \tag{1}$$

文章用公式(1)计算每一个特征和目标值的互信息值，通过大小顺序排列，选择互信息值较大的特征。

3.2. XGBoost 算法

XGBoost [13]由华盛顿大学陈天奇博士提出，它是 Gradient Boosting 实现的有监督学习算法，可以解决分类、回归等问题。训练采用的数据集样本为(x_i,y_i)，其中$x_i\in R^m$，$y_i\in R$。x_i表示具有 m 维的特征向量，y_i表示样本标签，模型包含 K 棵树，则 XGBoost 模型的定义如下：

$$\hat{y}_i=F_K(x_i)=F_{K-1}(x_i)+f_K(x_i) \tag{2}$$

$f_K(x)$表示第 K 棵决策树，决策树会对样本特征进行映射，使每个样本落在该树的某个叶子节点上。每个叶子节点均包含一个权重分数，作为落在此叶子节点的样本在该树的预测值ω。计算样本在每棵树的预测值(即ω)之和，并将其作为样本的最终预测值。

XGBoost 的目标函数定义如下：

$$\text{Obj}=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{3}$$

目标函数 Obj 有两项组成：第一项为损失函数，用于评估模型预测值和真实值之间的损失和误差，该函数必须是可微分的凸函数；$\Omega(f)$项为正则项，用来控制模型的复杂度，正则化项倾向于选择简单的模型，避免过拟合。正则化项的定义如下：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2 \tag{4}$$

第一项γT通过叶子节点树及其系数控制数的复杂度，值越大则目标函数越大，从而抑制模型的复杂程度。第二项为 L2 正则项，用于控制叶子节点的权重分数。

为找到最优的$f(x_i)$使目标函数最优，对公式中的目标函数，XGBoost 采用了近似的方法。对式(3)改写：

$$\text{Obj}^{(s)}=\sum_{i=1}^{n}L\left(y_i,\hat{y}_i^{(s-1)}+f_s(x_i)\right)+\Omega(f_s) \tag{5}$$

式(5)中，$\hat{y}_i^{(s-1)}$为第$s-1$轮样本x_i的模型预测值，$f_s(x_i)$为第 s 轮训练的新子模型。

XGBoost 引入泰勒公式近似和简化目标函数。取二阶泰勒公式的定义如下：

$$f\left(x+\Delta x\right)\cong f\left(x\right)+f'\left(x\right)\Delta x+\frac{1}{2}f'\left(x\right)\Delta x^2 \tag{6}$$

将式(6)中的 $\hat{y}_i^{(s-1)}$ 看作 x,将 $f_s\left(x_i\right)$ 看作 Δx，对 XGBoost 目标函数进行泰勒展开

$$\text{Obj}^{(s)}=\sum_{i=1}^{n}\left[L\left(y_i,\hat{y}_i^{(s-1)}\right)+g_i f_s\left(x_i\right)+\frac{1}{2}h_i f_s^2\left(x_i\right)\right]+\Omega\left(f_s\right) \tag{7}$$

式(7)中 g_i 为损失函数的一阶梯度统计；h_i 为二阶梯度统计。g_i、h_i 分别如下：

$$g_i=\frac{\partial L\left(y_i,\hat{y}^{(s-1)}\right)}{\partial\hat{y}^{(s-1)}}, \tag{8}$$

$$h_i=\frac{\partial L^2\left(y_i,\hat{y}^{(s-1)}\right)}{\partial\hat{y}^{(s-1)}} \tag{9}$$

去掉常数项 $L\left(y_i,\hat{y}^{(s-1)}\right)$，并将 $\Omega\left(f_s\right)$ 表达式代入公式，则式(7)转换为

$$\text{Obj}^{(s)}=\sum_{i=1}^{n}\left[g_i f_s\left(x_i\right)+\frac{1}{2}h_i f_s^2\left(x_i\right)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{10}$$

3.3. 随机森林

随机森林[14]是一种集成学习方法，由美国科学家 Leo Breiman 在 2001 年提出的一种机器学习方法。该方法结合 Bagging 集成学习理论和随机子空间方法，将多个决策树作为基分类器，以一定策略选取部分属性和数据分别建树；在预测阶段根据森林中各棵树的预测结果进行投票表决，最终表决结果为随机森林预测结果。

3.4. 软投票策略

软投票策略[15]预测模型将所有模型预测样本以某一类别的概率的平均值作为标准，概率最高的对应的类型为最终的预测结果。将 XGBoost 和 RFC 两个算法通过软投票的方法进行预测，得到比单一算法更好的预测结果。

3.5. BalanceCascade 算法

BalanceCascade 算法是一个对于 EasyEnsemble 的算法的改进，它是借鉴了 AdaBoost 的纠错思想，选择更多不同的样本进行学习，从而求得更好的预测结果。

输入样本：小类样本集 P，大类样本集 N，其中 $|P|<|N|$；从 N 中采样的子集个数 T，迭代次数 s；训练 AdaBoost 集成学习器 H_i

输出：集成模型 $H\left(x\right)$

(a) $i=0$，$f=\sqrt[T-1]{\frac{P}{N}}$，其中 f 为误报率(即判定为小类的大样本占比)，训练 H_i；

(b) Repeat;

(c) 对 N 随机采样得到子集 N_i，其中 $N_i=P$；

(d) 由 P 和 N_i 训练 AdaBoost 集成学习器 H_i，其中弱学习器 $h_{i,j}$ 的个数为 s_i，相应权重为 $\alpha_{i,j}$。该集成学习器的阈值为 θ_i。公式如下：

$$H_i(x)=\operatorname{sgn}\left(\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x)-\theta_i\right) \tag{11}$$

(e) 调节阈值 θ_i 使得 H_i 的假正率为 f;

(f) 移除 N 中被 H_i 正确分类的全部样本;

(g) Until $i=T$;

(h) 输出集成学习器:

$$H_i(x)=\operatorname{sgn}\left(\sum_{i=1}^{T}\sum_{j=1}^{s_i}\alpha_{i,j}h_{i,j}(x)-\sum_{i=1}^{T}\theta_i\right) \tag{12}$$

4. 本文算法模型构建

预测用户购买主要分为三个步骤：1) 产生训练样本子集；2) 训练集分类器；3) 基分类器融合。

4.1. 产生训练样本子集

先将数据集分为购买和未购买两大类，再从这两类数据中进行随机抽样，使抽取的测试集数据中购买和未购买两类数据比例接近数据集中购买与未购买的数量比，再对抽取的数据进行特征提取，得到本文的训练样本子集。

4.2. 训练基分类器

随机森林是机器学习中常用的解决分类问题算法之一，是一个包含多个决策树的分类器。

XGBoost 是基于提升树模型设计的，而提升树被认为是性能最好的机器学习算法之一。此外，XGBoost 还支持二阶导数运算和并行计算，可以利用正则项控制模型复杂度，因此准确度高。

本文选择 XGBoost 和随机森林作为基分类器,由于 BalanceCascade 利用了集成策略，分别对每一组数据训练参数调优耗费时间过长，用集成后预测的结果来逐步调节算法的关键参数，从整体的角度进行调节参数，可以节约大量的时间，得到较优的结果。最后通过软投票方法将两种算法融合后做预测。

4.3. 算法融合方法

在本文的算法模型中运用了软投票的算法融合方法，在做对比实验中运用了经典的堆叠法(Stacking)做算法融合和当下较为流行的 XGBoost 和逻辑回归算法相结合的模型[16]。软投票方法在第二章进行了介绍，堆叠法是一种分层的集成模型，第一层用 XGBoost 和随机森林对训练样本做预测，第二层用逻辑回归算法将上一层的结果进行训练后再预测。XGBoost 和逻辑回归结合的方法最早用于互联网广告点击通过率预测，XGBoost 不仅是一个分类算法也是一个天然的特征处理器，通过样本落在 XGBoost 叶子节点上的位置构造一个新的特征，再通过逻辑回归算法做预测。

4.4. 基分类器的集成

通过前面的工作，可以得到 m 个不同的训练集，用软投票策略融合 XGBoost 和随机森林作为基分类器。运用 BalanceCascade 的方法将这些基分类器通过公式(12)联合起来做预测。

给定样本 $x=(x_1,x_2,\cdots,x_n)$， $x\in R^T$ ，集成模型有个基分类器 $h_1(x),h_2(x),\cdots,h_T(x)$ ，则对于样本 x，其集成输出的结果为

$$H(x)=\operatorname{sgn}\left(\sum_{i=1}^{T}h_i(x)-\frac{T}{2}\right) \tag{13}$$

4.5. 本文算法描述

算法 2：BalanceCascade-XGBoost&RFC

输入：未购买行为样本集 N 和发生购买行为样本集 P，且使得样本集 N 与 P 的比值关系满足数据集中未购买行为和发生购买行为的数据比；

基分类器的个数 T；

输出：购买行为预测结果 $H(x)$；

(a) 令 $i=1$，且分类器 $h_i=\text{softvoting}(\text{RFC},\text{XGBoost})$；

(b) 从样本集中 N 划分 m 份子样本集 $N_j\ (j=1,2,\cdots,m)$；

(c) 从样本集中 N 中除去 N_i 的数据后，对剩余的数据做不重复随机抽样，选取和样本集 P 相同数量的子样本集 D_i；

(d) 将样本集 D_i 分别与样本集 P 组合成训练集 train_datai，把训练过 train_datai 的分类器 h_i 去预测数据 N_i；

(e) 删除数据集 N 中分类器 h_i 预测数据集 N_i 的正确的数据，更新样本集 N；

(f) 用融合分类器 h_i 去预测样本集的结果，令 $i=i+1$，回到第(c)步重复执行，当 $i=T$ 时结束并执行下一步；

(g) 分别用 XGBoost 和随机森林去训练 T 个训练集 train_datai，并用软投票方法将 XGBoost 和随机森林结合后对测试集 test_data 预测得到 T 个结果；

(h) 用投票法将 T 个结果结合起来做出最终的二分类预测结果。

4.6. 本文算法优势

在处理样本类别不平衡问题上，本文采用 BalanceCascade 的采样方法，并与简单欠采样和 EasyEnsemble 的采样方法作对比。从算法上看，采用 BalanceCascade 采样方式，预测大类样本的准确率足够高时，不用刻意的的调节阈值。利用集成学习策略，从大类样本中随机抽样得到和小类样本数量相同的数据，在每一次迭代中，都要对大类样本的数据进行一次筛选，使能够选出更为不同的数据进行学习。而且 BalanceCascade 方法在对数据预测时候运用了 Bagging 的思想，能够更加充分的运用大类样本数据。同样在做预测的时候，本文选取的软投票策略将 XGBoost 和随机森林算法相融合。在该问题，XGBoost 和随机森林是预测效果很好的分类算法，但在预测的时候仍有些预测数据概率接近 0.5 产生预测失误，用软投票策略可以更好的降低这一误差，结果也证明如此。

5. 参数优化

5.1. XGBoost 调优

Max_depth、min_child_weight 和 colsample_bytree 是 XGBoost 的三个重要参数。Max_depth 为决策树的最大深度，本文将其设置为 1~32。Min_child_weight 表示叶节点的最小样本权重，本文将其取值范围设置为 1~6。Colsample_bytree 表示样本特征采样比，本文将其取值范围设为 0.5~1。通过对这些参数的逐一搜索，得到的 XGBoost 算法的最优参数选择如下:max_depth 为 12,min_child_weight 为 2,colsample_bytree 为 0.68。

5.2. 随机森林调优

N_estimators、max_depth、min_child_leaf 是随机森林的三个重要参数。N_estimators 是决策树的数量，

本文将其取值范围设为 10~200，其中 max_depth 为决策树的最大深度，本文将其取值范围设置为 1~40。Min_child_leaf 表示最小的叶节点样本数，取值范围设置为 1~50。通过对这些参数逐一搜索，得到随机森林的最优参数选择如下：n_estimators 为 140,max_depth 为 36,min_samples_leaf 为 8。

6. 实验及结果分析

6.1. 实验数据集

本文实验数据来源于阿里巴巴天池大赛提供的 2014 年 11 月 18 日到 2014 年 12 月 18 日本文选用天池大赛中阿里巴巴移动电商平台数据集进行测试，包括 11 月 18 日到 12 月 18 日的 2,084,859 条用户历史购买行为数据：包含用户标识、商品标识等 6 条字段，10,000 名用户，1054 种类的 422,858 件商品，对 11 月 27 日，28 日和 29 日数据用户对商品的购买行为作为实验数据。其中，行为信息有浏览、收藏、加购物车、购买四种方式。

6.2. 数据预处理

对于提供的数据，将重复的数据去掉，再用本文给出的采样方法来抽取数据，通过特征工程对抽取的数据生成衍生特征，其中产生了许多空数据，为了能够更好的处理数据，对空数据进行了填充。文章并没有用传统的方法填充数据，而是根据文章产生空数据的特点，如在提取购买浏览转化率的特征时，用户没有浏览商品，则购买浏览转化率的空值意义更接近于无穷大，所以文章用远大于其它特征值的数字进行填充。购买收藏转化率和购买购物车的转化率都是用同样的方法进行填充。得到了新的 108 维特征，对数据进行归一化处理后，通过互信息方法对数据进行特征选择，选择了在前 58 位的特征值。

6.3. 评价指标

文章采用精确率 P、召回率 R 和 F1 值三个指标来对模型的性能进行评估。其中文章将预测类别组合化为真正例(TP)、假正例(TN)、真负例(FP)、假负例(TN)，计算公式为

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

6.4. 实验设计

为了验证本文提出的预测模型及算法的有效性，实验使用上述的数据集和评价标准，用软投票的方法将预测结果较好的算法进行融合并与当前热门的机器学习模型和不同的算法融合模型进行纵向对比。再对数据集进行不同采样方法的横向对比，其中包括简单欠采样，EasyEnsemble 和 BalanceCascade 三种种采样方法。最后文章对特征选择前和特征选择后的数据结果进行对比。

6.5. 实验结果

以查准率 P、召回率 R 和 F1 值为评价指标，当前热门的机器学习方法进行纵向对比，与不同的采样方法进行横向对比，特征选择前的表 5 和特征选择后的表 6 进行对比。

Table 5. Comparison of different algorithms before feature selection
表 5. 特征选择前的不同算法对比

模型	简单欠采样			EasyEnsemble			BalanceCascade		
	P	R	F1	P	R	F1	P	R	F1
LR	0.443	0.717	0.548	0.415	0.721	0.550	0.693	0.658	0.675
SVM	0.428	0.726	0.539	0.412	0.734	0.527	0.753	0.646	0.695
GBDT	0.592	0.719	0.649	0.611	0.717	0.659	0.708	0.707	0.708
RF	0.602	0.734	0.661	0.622	0.731	0.672	0.733	0.712	0.722
XGBoost	0.593	0.742	0.660	0.656	0.754	0.701	0.706	0.733	0.719
RF&XGBoost	0.689	0.749	0.672	0.756	0.757	0.703	0.733	0.736	0.735
GBDT&XGBoost	0.619	0.751	0.679	0.635	0.757	0.691	0.722	0.720	0.721
GBDT&RF	0.602	0.735	0.662	0.641	0.744	0.688	0.740	0.707	0.723
Stacking	0.618	0.736	0.672	0.658	0.762	0.706	0.728	0.725	0.727
XGBoost + LR	0.650	0.671	0.660	0.729	0.699	0.713	0.736	0.693	0.714

Table 6. Comparison of different algorithms after feature selection
表 6. 特征选择后的不同算法对比

模型	简单欠采样			EasyEnsemble			BalanceCascade		
	P	R	F1	P	R	F1	P	R	F1
LR	0.411	0.728	0.525	0.413	0.723	0.525	0.710	0.653	0.681
SVM	0.431	0.709	0.536	0.412	0.703	0.520	0.750	0.627	0.683
GBDT	0.598	0.722	0.654	0.622	0.717	0.666	0.726	0.696	0.710
RF	0.606	0.716	0.656	0.655	0.722	0.663	0.733	0.703	0.717
XGBoost	0.579	0.742	0.651	0.662	0.772	0.712	0.718	0.730	0.724
RF&XGBoost	0.629	0.750	0.684	0.652	0.759	0.701	0.731	0.720	0.725
GBDT&XGBoost	0.624	0.742	0.678	0.637	0.750	0.689	0.733	0.715	0.724
GBDT&RF	0.626	0.719	0.669	0.629	0.726	0.674	0.733	0.694	0.713
Stacking	0.601	0.731	0.660	0.652	0.751	0.698	0.728	0.721	0.724
XGBoost + LR	0.626	0.681	0.652	0.732	0.695	0.713	0.739	0.702	0.720

(a) 对比实验中，本文所提出的基于 BalanceCascade 用软投票方法将 XGBoost 和随机森林融合的方法做预测的 F1 值最高，则对用户购买行为预测效果最好；

(b) BalanceCascade 方法比 EasyEnsemble 和简单欠采样方法预测结果更好，说明 BalanceCascade 对于平衡样本和数据训练的效果更好；

(c) 进行特征选择后的预测结果与特征选择前的预测结果相差不大，甚至在一些算法的预测结果还略优于特征选择前的结果，说明特征选择不仅能够减少计算量，还能减少一些带有噪声的数据；

(d) 基于 BalanceCascade 的软投票策略的方法的分类效果比其它的模型要优秀很多，主要是因为 BalanceCascade 的采样方法能够使大类样本的数据更加的充分的利用，而且在一些方面也降低了模型的偏差和方差。使用软投票的方法对 XGBoost 和随机森林算法融合得到的预测的结果比其它算法融合方法得到的预测结果更好，且也优于其它的单一算法。

7. 总结

本文通过分析阿里巴巴天池大赛所提供的数据，从用户，商品和用户－商品三个方面整理得到了108个不同特征，对数据信息进行了全面有效的挖掘。文章使用互信息的方法对数据进行特征选择，一方面是为了降低计算量，另一方面消除数据部分噪声。在处理类别不平衡问题方面，使用了BalanceCascade算法，并与EasyEnsemble算法和简单采样法作比较，它利用了集成学习机制，将从大类样本随机抽取和小类样本数量相同的数据进行学习，借鉴了自适应提升算法的思想，用软投票策略融合XGBoost和随机森林的算法作为基分类器进行预测，在每次迭代中删除与从大类样本随机抽取的数据属性相近的数据，使得下次迭代随机抽取的数据更加“与众不同”。并且使得全局不会丢失重要信息。本文章算法结合了Bagging和Boosting思想，降低了偏差和方差。通过用软投票的方法将XGBoost和随机森林进行融合的方法与单一的算法相比，它使得预测的结果概率更加均衡，降低了单一算法预测结果产生的偏差。在与热门的机器学习算法进行对比实验后，本文提出方法具有更高的F1值，也使得模型更加有效。

8. 展望

XGBoost和逻辑回归算法相结合的模型虽然在预测结果稍差于本文提出的算法模型，但是由于使用逻辑回归算法做预测，极大的缩短了算法的执行时间，从而提高了算法效率，且预测的准确率高于单一的逻辑回归算法，这说明XGBoost构造的特征比原始数据的特征更有助于做预测。实际生活中商品销售预测的数据量较大，在特征工程设计较好的情况下，运用这种方法效率更高。设计实用的特征工程则是该问题的一个研究方向。

基金项目

国家自然科学基金(12071293)；国家教育部人文社科规划基金项目(16YJA630037)；上海市一流学科建设项目(S1201YLXK)。

参考文献

[1] 李旭阳, 邵峰晶. LSTM与随机森林购买行为预测模型研究[J]. 青岛大学学报(工程技术版), 2018, 33(2): 17-20.

[2] 马倩. 基于机器学习的电子商务平台重复购买客户预测[D]: [硕士学位论文]. 兰州: 兰州大学, 2017.

[3] 陈龙. 基于机器学习方法的用户复购行为预测[D]: [硕士学位论文]. 天津: 南开大学, 2021.

[4] 张震. 基于机器学习算法的重复购买行为预测研究[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2019.

[5] 邹润. 基于模型组合算法的用户个性化推荐研究[D]: [硕士学位论文]. 南京: 南京大学, 2014.

[6] Tian, Y., Ye, Z., Yan, Y. and Sun, M. (2015) A Practical Model to Predict the Repeat Purchasing Pattern of Consumers in the C2C E-Commerce. *Electronic Commerce Research*, **15**, 571-583. https://doi.org/10.1007/s10660-015-9201-8

[7] Zuo, Y., Shawkat Ali, A.B.M. and Yada, K. (2014) Consumer Purchasing Behavior Extraction Using Statistical Learning Theory. *Procedia Computer Science*, **35**, 1464-1473. https://doi.org/10.1016/j.procs.2014.08.209

[8] Chang, H.J., Hung, L.P. and Ho, C.L. (2007) An Anticiaption Model Potential Customers' Pruchasing Behavior Based on Clustering Analysis and Association Rules Analysis. *Expert Systems with Applications*, **32**, 753-764. https://doi.org/10.1016/j.eswa.2006.01.049

[9] Cho, Y.S., Moon, S.C., Oh, I.B., Shin, J.-H. and Ryu, K.H. (2013) Incremenatal Weighted Mining Based on RFM Analysis for Rommending Prediction in U-Commerce. *International Journal of Smart Home*, **7**, 133-144. https://doi.org/10.14257/ijsh.2013.7.6.13

[10] Liu, X.-Y., Wu, J. and Zhou, Z. (2009) Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B, Cybernetics*, **39**, 539-550. https://doi.org/10.1109/TSMCB.2008.2007853

[11] Liu, T.Y. (2009) EasyEnsemble and Feature Selection for Imbalance Data Sets. *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, Shanghai, 3-5 August 2009, 517-520.

[12] 张尧. 基于互信息的特征选择方法研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2019.

[13] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[14] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[15] Macdonald, C. and Ounis, I. (2006) Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. *Proceedings of the* 2006 *ACM CIKM International Conference on Information and Knowledge Management*, Arlington, 6-11 November 2006, 387-396. https://doi.org/10.1145/1183614.1183671

[16] He, X., Pan, J., Ou, J., Xu, T., Liu, B., Xu, T., *et al*. (2014) Practical Lessons from Predicting Clicks on Ads at Facebook. *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, New York, 24-27 August 2014, 1-9. https://doi.org/10.1145/2648584.2648589