梯度渐进回归树算法在电子商务品牌推荐中的应用①

申端明, 乔德新, 许 琨, 林 霞, 江日念

(中国石油勘探开发研究院, 北京 100083)

摘 要: 针对电子商务推荐系统中, 互联网“信息过载”所造成的难以准确定位用户兴趣并提供准确品牌推荐的问题, 通过深入挖掘电子商务网中的用户行为日志, 抽取出能辨别出用户对商品品牌购买行为的多个特征, 然后将这些特征融入到梯度渐进回归树算法中, 建立用户兴趣偏好模型来提高推荐精度. 实验结果表明, 在数据稀疏的情况下, 该算法仍能较好的识别出用户对品牌的偏好, 并在推荐准确度方面较其他传统推荐和分类算法有明显的提高.

关键词: 品牌推荐; 梯度渐进回归树; 行为日志分析; 特征挖掘

Gradient Boosting Regression Tree Algorithm and Application of E-commerce Brand Recommendation

SHEN Duan-Ming, QIAO De-Xin, XU Kun, LIN Xia, JIANG Ri-Nian

(Research Institute, Petroleum Exploration & Development, Beijing 100083, China)

Abstract: In E-commerce recommendation system, “Information overload” on Internet has brought a tough problem, which is how to precisely position users’ interest and provide users with accurate brand recommendation. To solve this problem, in this paper, many features which could describe the purchasing behavior of users are extracted by deeply mining large-scale of user behavior logs. A brand preference model was constructed by applying these features into Gradient Boosting Regression Tree algorithm, to improve accuracy of the recommendation algorithm. Experiment results show that, in condition of sparse data, algorithm in this paper can still fit brand preference of users very well, and has significantly improvement in accuracy compared with traditional recommendation and classification algorithm.

Key words: brand recommendation; gradient boosting regression tree; behavior log analysis; feature mining

1 引言

伴随着互联网的迅速发展, 网络上记录的数据急速增长, 人们逐渐从信息匮乏的时代走入了信息过载的时代, 为了解决互联网的信息过载问题, 降低消费者的搜索成本, 推荐系统作为一种典型的信息过滤技术已广泛应用到电子商务系统中, 如 Amazon、MovieFinder、淘宝、当当等. 推荐算法模型是整个推荐系统的核心, 它的性能直接影响推荐效果, 因此推荐算法的研究已成为学术界和工业界共同关注的焦点问题.

在一些大型综合性购物网站, 每天都会有数千万的用户通过品牌发现自己喜欢的商品, 品牌是连接消费者与商品最重要的纽带. 好的品牌推荐方法可以起到保住老客户、吸引新客户的作用, 能有效提高电子商务网站的销售业绩, 提升用户的购物体验.

梯度渐进回归树[1](Gradient Boosting Regression Tree)又叫 MART(Multiple Additive Regression Tree), 是一种集成机器学习算法. 它通过集成多个弱决策树模型形成最终预测模型, 像其他 boosting 算法一样, 梯度渐进回归树也是通过级联方式构造最终模型, 该算法由多棵决策树组成, 所有树的结论累加起来做最终答案. 梯度渐进回归树算法具有评价特征重要性的能力, 对于噪声数据和存在缺失值的数据具有很好的鲁棒性, 它在被提出之初就和SVM一起被认为是泛化

① 收稿时间:2014-10-09;收到修改稿时间:2014-11-14

能力(generalization)较强的算法. 近些年更因为被用于搜索排序的机器学习模型而引起大家关注[2].

针对梯度渐进回归树的上述特点, 本文将梯度渐进回归树算法应用到电子商务网站的品牌推荐中, 根据用户在购物网站的行为日志, 提取用户、品牌和操作特征, 建立用户的品牌偏好, 然后基于这些特征应用梯度渐进回归树算法进行用户购买行为的预测研究. 最后, 本文在海量真实的天猫用户行为数据上进行了相应的实验, 实验结果表明, 梯度渐进回归树算法比其他传统推荐算法能更有效地预测用户对品牌下商品的购买行为, 从而提升推荐精度.

2 相关研究

现有的主要推荐方法可以分为三类: 基于关联规则的推荐算法[3]、基于内容的推荐算法[4]、协同过滤推荐算法[5].

基于关联规则的推荐算法是根据生成的关联规则模型和用户当前的购买行为为用户提供推荐服务. Agrawal 等[6]最早提出 Aprior 的关联规则推荐算法;为了提高 Aprior 的运行效率, Han 等[7]进一步提出了 FP-Growth 算法. 但是规则的发现极为耗时, 因此成为该算法最大的瓶颈, 并且随着规则的增加, 对系统的管理也变得越来越复杂.

基于内容的推荐方法对用户以前访问过的商品进行分析, 并将与其相似的未知商品推荐给用户, 这种方法主要是对商品的资料(如大小、类别、生产商等)进行分析, 然后将未知的商品与之比较以发现相似的商品[8]. 随着机器学习等技术的发展, 当前基于内容的推荐算法更多的是通过比较项目与用户描述文件来为用户提供推荐服务[9]. 但是该算法只能发现与用户历史兴趣相似的项目, 不能为用户发现新的感兴趣的资源.

协同过滤推荐算法的核心思想是: 根据用户行为分析用户兴趣, 在用户群中找到与目标用户(兴趣)相似的邻居用户, 综合这些邻居用户对某一信息的评价, 形成系统对该目标用户在商品喜好程度方面的预测, 系统再根据这些喜好程度进行相应的推荐. 协同过滤技术是目前推荐系统中应用最广泛及效果最好的技术之一[10]. 由 Goldberg 等[11]提出的 Tapestry 系统是最早的协同过滤系统. 该系统利用小型社区成员的直接观点来进行电子邮件分类过滤. 但是该系统不适合应用到大型社区中去, 所以各类型的协同过滤技术就陆续出现了. 例如 Sarwar 等[12]提出使用矩阵奇异值分解的方法降低评分矩阵维度以减少稀疏性. 邓爱林等[13]提出首先根据基于项目的协同过滤算法预测部分项目评分减少评分稀疏性再根据基于用户的协同过滤算法为用户进行推荐.

基于内容的推荐算法、协同过滤推荐算法在音乐推荐、视频推荐和新闻推荐等特定类型的商品推荐方面取得了显著的效果, 但用于品牌推荐时存在一定的不适应性, 这主要是由于品牌推荐和特定类型的商品推荐的性质不同决定的——需求的单一性与喜好的相似性. 用户喜欢看动作片电影, 那么相似的动作片电影都能看一遍; 用户看过某一类型的书籍, 那么相似类型的书籍可能都有兴趣. 但是购物就不同了, 购物更多的是刚需, 喜欢某一款的衣服, 一般人也不会把所有喜欢的这一款的衣服都买下来, 更多的是每个季节只买一件. 家居类的频次就低了, 可能一辈子就买一次, 不同的类目的需求频率是不一样的. 所以基于内容的推荐算法、协同过滤推荐算法并不适用于进行品牌推荐.

针对现有推荐方法对品牌推荐的不适应性, 本文根据用户在购物网站的行为日志提取特征, 引入梯度渐进回归树算法, 设计出了一个新的基于梯度渐进回归树的品牌推荐算法模型, 以提高推荐算法的精度.

3 原理与方法

3.1 梯度渐进回归树算法

Friedman[14]在 1999 年提出了 Gradient Boosting 算法, Gradient Boosting 与传统的 Adaboost 的区别在于, 再一次计算为了减少上一次模型的残差, 在残差减少的梯度方向上建立了一个新的模型, 由此不断迭代产生一个基础分类器的组合, 使得组合分类器可以对损失函数进行极小化优化. 因此 Gradient Boosting 算法能够在建模时, 使之前模型的残差往梯度方向减小, 与 Adaboost 对正确、错误的样本进行加权有很大的区别.

现给定数据样本$\{x_i, y_i\}_{i=1}^{n}$, 损失函数$L(y, F(x))$和基础分类器$\{h(x)\}$, 其中$x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$, p为预测变量的个数, y_i为预测标签, Gradient Boosting 算法的步骤过程如下:

1) 首先给定模型的初始值β(常数项):

$$F_0(x) = argmin_\beta \sum_{i=1}^{N} L(y_i, \beta) \quad (1)$$

2) 对于迭代次数$m = 1:M$(M为迭代次数), 求上次迭代模型$F_{m-1}(x)$的导数, 即求出残差的梯度方向:

$$\acute{y}_i = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}, \quad i=1,2,\ldots N \quad (2)$$

3) 把式(2)求得的值作为伪因变量, 用基础分类器$\{h(x)\}$拟合数据样本$\{x_i, \acute{y}_i\}_{i=1}^{n}$, 根据最小二乘原则, 求得模型参数$a_m$, 拟合的模型为$h(x_i, a_m)$.

4) 根据损失函数最小化原则, 求得模型新的步长β_m, 将β_m看做当前模型的权重:

$$\beta_m = \arg min \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i, a_m)) \quad (3)$$

5) 更新模型:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i, a_m) \quad (4)$$

6)M次迭代结束, 得到最终模型:

$$F(x) = \sum_{m=1}^{M} \beta_m h(x_i, a_m) \quad (5)$$

3.1.1 收缩系数

Gradient Boosting 算法引入了收缩参数, 收缩思想认为, 每次走一小步逐渐逼近结果的效果, 要比每次迈一大步很快逼近结果的方式更容易避免过拟合. 即它不完全信任每一个棵残差树, 它认为每棵树只学到了真理的一小部分, 累加的时候只累加一小部分, 通过多学几棵树弥补不足. 用方程表示原模型更新为:

$$F_m(x) = F_{m-1}(x) + v \cdot \beta_m h(x_i, a_m), 0 < v \leq 1 \quad (6)$$

参数v称为"学习率", 应用较小的学习率($v < 0.1$)无收缩参数的 Gradient Boosting ($v = 1$) 算法更能有效提高模型的泛化能力. 虽然学习率越小, 学习越充分, 但较小的学习率也需要更大的迭代次数才能使模型收敛, 这就导致了计算时间的增加, 所以学习率参数的选择也需要一定的权衡.

3.1.2 随机化参数

Friedman 结合 Breiman 的 bagging 思想, 在 Gradient Boosting 算法基础上引入了随机化参数, 提出了 SGB 算法. 其主要思想是在算法的每次迭代过程中, 随机抽取训练样本的一部分来拟合基础分类器, 抽样比例由随机化参数 f 表示. 当 f=1 时, 算法等同于原始 Gradient Boosting 算法, 较小的 f 值表示训练的样本更少, 计算耗时更短, 但能够有效地避免过拟合, 当 f=0.5 时, 模型泛化能力最佳.

3.2 特征提取

在梯度渐进回归树算法预测中, 选择恰当的特征是极为重要的一步, 如果选择了具有充分辨别能力的特征, 将极大地提高决策树分类器的精度.

在用户的行为日志中主要包括用户标记、行为时间、用户对品牌的行为类型和品牌标记四个字段, 如表 1; 通过对用户日志分析, 将购买预测(品牌推荐)问题转化为三个简单问题: 1)哪些用户会购买, 2)哪些品牌会被购买, 3)用户会购买哪些品牌; 主要提取以下三类特征: 用户特征, 品牌特征, 用户--品牌特征.

表 1　用户行为日志字段说明

字 段	字段说明	提取说明
user_id	用户标记	抽样&字段加密
Time	行为时间	精度到天级别&隐藏年份
action_type	用户对品牌的行为类型	包括点击、购买、加入购物车、收藏 4 种行为(点击: 0 购买: 1 收藏: 2 购物车: 3)
brand_id	品牌数字 ID	抽样&字段加密

3.2.1 用户特征

用户特征的主要目的是识别哪些用户可能会产生购买行为. 主要特征包括商品类、品牌类、购物热度类和用户对品牌偏好 4 类、共计 15 个特征, 如表 2 所示.

表 2　用户类特征说明

类别	特征名称	特征描述
商品类	购买商品量	用户购买商品量
	点击商品量	用户点击商品量
	收藏商品量	用户收藏商品量
	购物车商品量	用户添加购物车商品量
品牌类	购买品牌量	用户购买品牌量
	点击品牌量	用户点击品牌量
	收藏品牌量	用户收藏品牌量
	购物车品牌量	用户添加购物车品牌量
购物热度	点击到购买商品的转化率	购买商品量/点击商品量
	收藏到购买商品的转化率	购买商品量/收藏商品量
	购物车到购买商品的转化率	购买商品量/购物车商品量
	平均购买量	购买商品量/活跃天数
用户对品牌偏好	点击到购买品牌的转化率	购买品牌量/点击品牌量
	收藏到购买品牌的转化率	购买品牌量/收藏品牌量
	购物车到购买品牌的转化率	购买品牌量/购物车品牌量

3.2.2 品牌特征

品牌特征的主要目的是识别哪些品牌商品可能会被购买. 主要特征包括品牌操作类、品牌购买转化率类、品牌热度类 3 类、共计 21 个特征, 如表 3 所示.

表 3 品牌类特征说明

类别	特征名称	特征描述
品牌操作类	品牌销售量	品牌被购买次数
	品牌点击量	品牌被点击次数
	品牌收藏量	品牌被收藏次数
	品牌购物车量	品牌被加入购物车次数
	品牌购买人数	购买品牌的人数
	品牌点击人数	点击品牌的人数
	品牌收藏人数	收藏品牌的人数
	品牌购物车人数	加入品牌到购物车的人数
	人均点击数	品牌点击量/品牌点击人数
	人均购买量	品牌销售量/品牌购买人数
	人均收藏量	品牌收藏量/品牌收藏人数
	人均购物车量	品牌购物车量/品牌购物车人数
购买转化率类	品牌点击到购买转化率	品牌销售量/品牌点击量
	品牌收藏到购买转化率	品牌销售量/品牌收藏量
	品牌购物车到购买转化率	品牌销售量/品牌购物车量)
	品牌点击到购买人次转化率	品牌购买人数/品牌点击人数
	品牌收藏到购买人次转化率	品牌购买人数/品牌收藏人数
	品牌购物车到购买人次转化率	品牌购买人数/品牌购物车人数
品牌热度类	品牌活跃度	有3次及以上点击的用户数/总用户数
	品牌跳出率	只对品牌进行过 1 次点击操作的用户数/总用户数
	品牌返客率	多次购买的用户数/总购买用户数

3.2.3 用户-品牌特征

用户--品牌类特征的主要目的是识别用户是否会对特定品牌产生购买行为. 主要特征包括用户对品牌操作类、计算类、综合类 3 类、共计 22 个特征、如表 4 所示.

表 4 用户–品牌类特征说明

类别	特征名称	特征描述
用户对品牌操作类	访问量	用户对品牌的行为次数
	点击量	用户对品牌的点击次数
	购买量	用户对品牌的购买次数
	收藏量	用户对品牌的收藏次数
	加入购物车量	用户对品牌的加入购物车次数
	点击天次	一天内多次点击去重
	购买天次	一天内多次购买去重
	收藏天次	一天内多次收藏去重
	购物车天次	一天内多次加入购物车去重
	最后一次点击时间	用户对品牌的最后一次点击时间
	最后一次购买时间	用户对品牌的最后一次购买时间
	最后一次收藏时间	用户对品牌的最后一次收藏时间
	最后一次加入购物车时间	用户对品牌的最后一次加入购物车时间
计算类	点击天数的一阶均值	点击天数的一阶均值
	点击天数的二阶方差	点击天数的二阶方差
	点击天数的三阶峰度	点击天数的三阶峰度
	购买天数的一阶均值	购买天数的一阶均值
	购买天数的二阶方差	购买天数的二阶方差
	购买天数的三阶峰度	购买天数的三阶峰度
综合类	购买率	购买量/用户购买商品量
	点击率	点击量/用户点击商品量
	偏好评分	不同操作权重分数不同

3.3 算法实现

根据用户行为日志, 提取计算出上述多种特征; 通过用户在预测时间段内的行为日志标定用户购买行为, 构建出训练数据集(1 表示购买, 为正样本, 0 表示非购买, 为负样本).

模型输入:

训练数据集 S,预测数据集 T,决策树个数 k, 以及特征个数 M

模型训练:

① 给定模型初始值 $F_0(0)=0$,

② 基础分类器 $h(x)$为一个回归树分类器, 分类原则为在每次分枝时, 穷举每一个特征的每个阈值, 找最好的分割点, 衡量最好分割点的标准是每个样本的预测误差平方和除以样本个数. 分枝直到每个叶子节点上样本的标定值都唯一或者达到预设的终止条件(如叶子个数上限), 若最终叶子节点上样本的标定值不唯一, 则以该节点上所有样本的标定值的平均值作为该叶子节点的预测值.

③ 对于迭代次数 m=1:M(M 为迭代次数), 通过随机化参数 f=0.5, 随机抽取出新的训练数据集 S_i,在新的训练数据集 S_i 上, 求上次迭代模型 $F_{m-1}(x)$的导数, 即求出残差的梯度方向, 在残差减少的梯度方向上应用回归树分类器建立了一个新的决策树模型, 由此不断产生一个基于基础分类器的级联组合, 使每一个新的决策树模型都能在上一个模型的基础上更接近真实值. 通过引入压缩系数, 每一个新的模型更缓慢的逐渐逼近结果值, 从而提高模型的泛化能力.

④ 循环进行第 3 步, 通过 M 次迭代构建出梯度渐进回归树模型.

模型预测:

利用以构建成的对预测数据集 T 进行预测, 梯度渐进回归树模型的 M 个决策树模型将依次对每个样本的目标变量进行预测, 最后累加形成最终的预测结果.

模型输出:

预测数据集 T 中每个人对商品的购买行为值(介于 0-1 之间, 越接近 1, 购买的可能性越大).

4 实验结果与分析

4.1 实验数据集

本文使用天猫海量真实用户的访问日志数据作为实验数据集. 天猫是亚洲最大的综合性购物平台, 拥有 10 万多品牌商家, 每天有上千万的用户在该平台进行访问、购物活动. 本文数据集来自阿里巴巴天池平台, 天池平台是阿里巴巴自主研发的分布式计算平台, 其中涵盖了 MapReduce、SQL 编程及各种平台集成的机器学习算法包, 适用于大数据处理与建模. 本文抽取了天猫用户日志中的四个字段, 数据样例见如表 5. 该数据集包含 12436614 个用户对 29552 个品牌的 4 个月份的点击、购买、收藏、加入购物车四种行为数据记录. 用户对任意商品的行为都会映射为一行数据, 其中所有商品 ID 都已汇总为商品对应的品牌 ID, 用户和品牌都分别做了一定程度的数据抽样, 且数字 ID 都做了加密, 所有行为的时间都精确到天级别(隐藏年份).

表 5　实验数据集样例

user_id	brand_id	type	visit_datetime
12154500	1758	0	4/15
12154500	1758	0	4/15
5780000	18736	1	4/15
364250	11679	2	4/16
5078750	14580	3	4/18
2088000	25079	3	4/22

将实验数据集中最后一月的 2442730 条用户购买行为数据划分为验证数据集, 剩余三个月的数据划分为测试数据集, 最后将剩余 3 个月的数据按 3:1 比例进行标注处理得到训练数据集.

评价标准

通过前三个月的用户对品牌的行为记录, 预测第四个月的用户对品牌下商品的购买行为. 评价原则为: 预测的品牌准确率越高越好, 覆盖的用户和品牌越多越好. 具体采用最常用的准确率与召回率作为两个主要评价指标.

准确率: $$\text{Precision}=\frac{\sum_i^N hitBrands_i}{\sum^N pBrands_i} \quad (7)$$

其中N为预测的用户数, $pBrands_i$为对用户i预测的品牌个数, $hitBrands_i$对用户i预测的品牌与用户i真实购买的品牌交集的个数;

召回率: $$\text{Recall}=\frac{\sum_i^M hitBrands_i}{\sum^M bBrands_i} \quad (8)$$

其中 M 为实际产生成交的用户数, $bBrands_i$为用户i真实购买的品牌个数, $hitBrands_i$对用户i预测的品牌与用户i真实购买的品牌交集的个数;

最后用 F1-Score 来你和准确率和召回率, 并以 F1-Score 作为最终的评价指标.

$$F_1=\frac{2*P*R}{P+R} \quad (9)$$

4.2 实验设计

在对实验数据分析中发现, 用户行为操作时间距离预测时间越近对预测结果的准确性影响也越大, 为此, 本文将训练数据集中的数据分为前 3 天、7 天、15 天和剩余天数四个区间, 原来的 58 个特征增加为 216 个特征. 在构建成功的训练数据集中, 正负样本的比例约为 1:400, 正样本在数据表中相对是比较稀疏的, 针对此问题, 应不断调整正负样本的比例, 设计多组实验进行验证. 在梯度渐进回归树算法中决策树的个数对预测结果也起着至关重要的作用. 由于正负样本比例和决策树个数都对实验结果产生影响, 为简化实验过程, 本实验在评估一个参数之前固定另一个参数.

图 1 给出了决策树个数 k=30, 正负样本比例与预测结果 $F1$ 之间的曲线关系. 由图可知, 当正负样本比例为 1:6 时, 实验结果 $F1$ 分数最高.

图 2 给出了正负样本比例为 1:6 时, 随机森林决策树个数与预测结果 $F1$ 之间的曲线关系. 由图可知, 随着决策树个数 k 的增多, $F1$ 分数逐渐增加, 但增加趋势不同, 当 k 在[10,30]区间时, $F1$ 分数上升趋势明显, 当 k>30 时, 上升趋势逐渐下降, 并且当 k>100 后, $F1$ 分数逐渐成平稳状态. 然而, 当随着决策树个数的增多, 模型预测时间也是呈直线上升趋势的, 综合考虑 $F1$ 分数和时间因素, k=150 为最优.

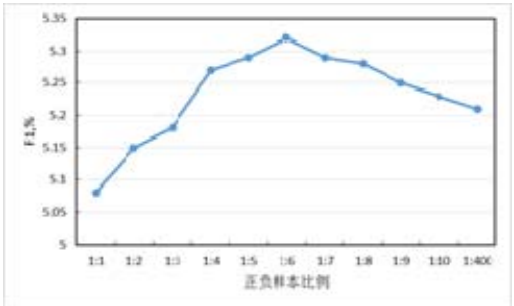

图 1　正负样本比例与 $F1$ 分数

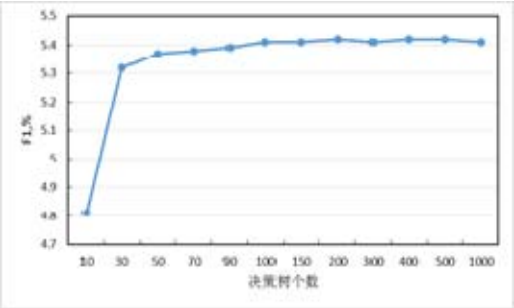

图 2 决策树个数与 *F*1 分数

图 1 给出了决策树个数 *k*=30, 正负样本比例与预测结果 *F*1 之间的曲线关系. 由图可知, 当正负样本比例为 1:6 时, 实验结果 *F*1 分数最高.

图 2 给出了正负样本比例为 1:6 时, 随机森林决策树个数与预测结果 *F*1 之间的曲线关系. 由图可知, 随着决策树个数 *k* 的增多, *F*1 分数逐渐增加, 但增加趋势不同, 当 *k* 在[10,30]区间时, *F*1 分数上升趋势明显, 当 *k*>30 时, 上升趋势逐渐下降, 并且当 *k*>100 后, *F*1 分数逐渐成平稳状态. 然而, 当随着决策树个数的增多, 模型预测时间也是呈直线上升趋势的, 综合考虑 *F*1 分数和时间因素,*k*=150 为最优.

4.3 算法对比

针对梯度渐进回归树算法(GBRT), 本节将与逻辑回归(LR)、支持向量机(SVM)、品牌评分算法(BG)和随机森林算法进行比较. 由表 3 可见, 不同算法效果相差明显. RFBR 算法 *F*1 分数达到 5.41%, 相比于 RF 的 5.33%, 提高了 0.08%, 其他算法表现更加一般, SVM 效果最差, 仅为 1.43%. 相对于其他算法, GBRT 算法和 RF 算法具有如此优越的表现主要是这两种算法中的多个决策树构成的强分类器相对其他算法中的弱分类器的优势. 同时 GBRT 算法也具有较强的泛化能力和模型健壮性, 相对于随机森林算法的多个分类树组合, 梯度渐进回归树算法通过迭代, 使每一棵树学习的是之前所有树结论和的残差, 在本实验结果中往往比随机森林算法效果更好.

表 3 品牌推荐算法效果比较

算法	Precision(%)	Recall(%)	F_score(%)
LR	3.13	3.27	1.20
SVM	1.43	1.29	1.36
BG	3.42	3.41	3.41
RF	5.33	5.33	5.33
GBRT	5.41	5.41	5.41

5 结语

本文充分分析用户在购物网站的行为日志, 从中抽取出能辨别出用户对商品品牌购买行为的多个特征, 然后将这些特征融入到梯度渐进回归树算法中, 从而提高品牌推荐预测的精度. 由于用户的行为日志比较容易获得, 因此基于梯度渐进回归树的品牌推荐算法可应用的范围更加广泛. 同时由于随机森林对于噪声数据和存在缺失值的数据具有很好的鲁棒性, 基于梯度渐进回归树的品牌推荐算法也具有稳定性好、可扩展性强等特点. 在真实的天猫用户访问数据集上的实验表明, 该方法针对传统的推荐和分类算法相比具有明显的效果提高.

参考文献

1 Friedman JH. Greedy function approximation: a gardient boosting machine. Annal of Statistics, 2001, (29): 1189–1232.

2 章光明,刘晋,贾慧珣,等.随机梯度 boosting 算法在代谢组学研究中的应用.中国卫生统计,2013(03):323–326.

3 Panagiotiss, Alexandrosn, Apstolosn, et al. Collaborative recommender system: combing effectiveness and efficiency. Expert Systems With Applications, 2007, 34(4): 2995–3013.

4 Sung-Shun W, Lin BS, Wen-Tien C. Using contextual Information and multidimensional approach for recommend-dation. Expert Systems With Applications, 2009, 36(2): 1268–1279.

5 Leungcw, Chansc, Chungf, et al. An empirical study of a cross–level association rule mining approach to cold-start recommendations. Knowledge-Based Systems, 2008, 21(7): 515–529.

6 Agrawal R, Imuekubsju R, Swami A. Mining association rules between sets of items in large databases. Proc. of ACM SIGMOD International Conference on Management of Data. New York. ACM Press. 1993. 207–216.

7 Han JW, Pei J, Yin YW, et al. Mining frequent patterns without candidate Generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2004, 8(1): 53–87.

8 刘枚莲,刘同存,李小龙.基于用户兴趣特征提取的推荐算法研究.计算机应用研究,2011,(5):1664–1667.

9 Liu DR, Shih YY. Hybrid approaches to product recommendation base on customer lifetime value and

purchase preferences. Journal of Systems and Software, 2005, 77(2): 181–191.

10 扈中凯,郑小林,吴亚峰,等.基于用户评论挖掘的产品推荐算法.浙江大学学报(工学版),2013(8):1475–1485.

11 Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12): 61–70.

12 Sarwar B, Karypis G, Konstan J, et al. Application of dimensionality reduction in recommender system-a case study. Minnesota Univ Minneapolis Dept of Computer Science, 2000.

13 邓爱林,左子叶,朱扬勇,等.一种改进的基于项目聚类的协同过滤推荐算法.小型微型计算机系统,2004,25(9):1665–1670.

14 Friedman JH. Stochastic gradient boosting. Computatianal Statistics & Data Analysis, 2002, (38): 36.