基于 XGBoost 算法的用户评分预测模型及应用*

杨贵军[1]　徐　雪[1]　赵富强[2]
[1](天津财经大学中国经济统计研究中心　天津 300222)
[2](天津财经大学理工学院　天津 300222)

摘要: 【**目的**】基于用户网络评论构建有效的评分预测模型，挖掘用户消费行为特征。【**方法**】基于 LDA 模型，量化用户评论为主题特征向量作为解释变量，将用户评分作为被解释变量，采用 XGBoost 算法，并加入样本扰动和属性扰动生成多个模型进行集成，构建用户评分预测模型。【**结果**】针对某汽车门户网站的用户评论评分预测结果表明，该模型较好地揭示了用户对汽车商品的偏好。较逻辑回归、随机森林算法，其预测准确度分别高出 13.73%、0.64%，且具有较高的计算效率。【**局限**】未融合其他方面的数据对用户行为特征进行更全面的刻画。【**结论**】将用户评论量化为主题特征向量，基于 XGBoost 算法能够准确、高效地预测用户评分。

关键词: 评分预测　XGBoost 算法　LDA 主题模型　文本特征提取　用户评论
分类号: G35
DOI: 10.11925/infotech.2096-3467.2018.0414

1　引　言

随着互联网和社交网络的快速发展，越来越多的厂商和消费者开始关注商品的用户评价。用户评价既是消费者消费决策的参考，也是生产厂商掌握消费者偏好、优化生产行为的重要依据。

商品的用户评价是指消费者对所购买商品属性特征的主观感受，反映消费者对购买商品的偏好和满意度，包括用户对商品的评分信息和文本评论信息。其中，用户评分是消费者对商品属性特征以及使用满意度的综合评价，消费者和厂商通常会优先关注评分较高的商品；商品评论中蕴含着描述商品属性特征和用户满意度的关键词等信息。鉴于用户评论和评分之间存在关联性，如何将用户评论与评分有效结合，构建预测准确度高、时效快的评分预测模型，以预测用户消费偏好，为产品研发决策和商品推荐提供参考是一个重要的研究课题。

基于此，本文在回顾已有方法的基础上，结合用户评论与评分信息，构建一种基于 XGBoost 算法的用户评分预测模型。利用 LDA(Latent Dirichlet Allocation)模型提取用户评论的主题特征，并依据主题特征及其概率分布将文本评论信息量化成主题特征向量；针对主题特征向量与用户评分，加入数据样本扰动和属性扰动建立多个不同的 XGBoost(eXtreme Gradient Boosting)模型，采用投票法进行集成，提高模型的泛化能力。通过改善用户评分预测准确性，提高计算效率，以分析用户评论主题与评分之间的相关性，挖掘用户消费偏好。

2　国内外相关研究

评分预测问题是推荐系统研究的一个分支，评分

通讯作者：徐雪，ORCID: 0000-0002-3421-1565，E-mail: xuxue2017@163.com。

*本文系国家自然科学基金面上项目“劣者淘汰两阶段自适应临床试验的设计和分析”(项目编号: 11471239)、国家社会科学基金青年项目“社交媒体中敏感信息可信度评估方法研究”(项目编号: 18CTJ008)和全国统计科研计划重点项目“Web 社会网络中敏感信息识别及突发事件预测研究”(项目编号: 2017LZ05)的研究成果之一。

预测的准确性将很大程度上影响推荐系统的性能，很多学者针对评分预测问题展开研究。早期关于评分预测的研究，多是基于用户历史评分行为和物品属性特征进行建模[1-3]。随着互联网和电子商务的快速发展，用户参与度不断提高并生成大量的评论信息。相对于评分和物品属性特征而言，用户评论信息蕴含的内容更为丰富，能够更加具体、准确地表达用户对物品的喜好，这为评分预测模型的构建提供了新研究思路。研究人员开始关注将评分与评论相结合的方法，从评论文本中挖掘用户偏好提高评分预测的质量。Li 等[4]通过手动建立部分主题词，将评论文本主题与评分矩阵分解模型融合，估计用户在不同方面的偏好。Fan 等[5]从 Yelp 餐馆评论数据中提取高频词和高频形容并词创建词袋，通过与评分数据相结合，利用线性回归模型实现评分预测。张红丽等[6]针对电影评论数据，提取用户语料中的情感特征作为辅助预测指标，并结合评论人数等与评论相关联的指标作为自变量，利用回归分析构建评分预测模型。高祎璠等[7]基于餐馆评论数据，利用用户评论的主题分布建立用户画像和商品画像，基于逻辑回归建立评分预测模型，预测准确度为 52%，在个别子数据集上预测准确度可达 66%。虽然现有方法在一定程度上利用用户评论信息，构建的用户评分预测模型仍面临准确率较低的问题，如何将用户评论与评分有效结合，构建准确、高效的用户评分预测模型，这项工作还有待深入研究。

在实际应用中，用户评分往往会依据分值大小进行分类，将评分预测问题转化为二分类或多分类问题。分类模型的选择将会影响评分预测模型的性能。常用的分类模型有朴素贝叶斯、逻辑回归、随机森林等。其中，朴素贝叶斯分类方法假设样本的各个属性相互独立，在实际问题中被广泛应用，但如果用户评分之间存在相互依赖关系时，将会影响算法的准确性。线性回归方法是基于用户历史评分的线性推荐算法，线性回归模型的参数估计基于用户–项目矩阵，当矩阵存在稀疏、噪声等问题时，方法准确率可能会降低[8]。

现有研究结果显示，相对于逻辑回归、决策树等单一分类器，根据训练数据构建一组个体学习器，并采用某种策略将多个学习器进行集成的学习方法具有更高的准确度和更好的稳健性[9-10]。集成学习方法主要分为两类: Bagging 方法(如 RF 算法等); Boosting 方法(如 XGBoost 算法等)。其中，RF[11]算法利用样本扰动和属性扰动实现基学习器的多样性，提升算法的泛化性能，但该算法需要存储每棵决策树及其每个节点不同的样本集合，内存开销较大，模型训练速度较慢。XGBoost[12]算法依据损失函数在梯度下降方向上组合多个 CART 树，以最小化损失函数，且能够自动利用 CPU 的多线程进行分布式学习和多核计算，在保障分类准确度的前提下可以提高计算效率，适用于处理大规模数据。近年来，XGBoost 算法在文本数据中的应用也开始受到关注，取得较好的效果[13-14]。鉴于此，本文在模型底层使用 XGBoost 算法进行 Boosting 集成，以提高算法效率；在顶层，为了增加集成学习中个体学习器的多样性，利用 Bagging 的思想加入数据样本扰动和属性扰动，以建立多个好而不同的 XGBoost 模型，并采用投票法进行集成，提高模型的泛化能力。

3　基于 XGBoost 算法的用户评分预测模型

基于 XGBoost 算法的用户评分预测模型主要包括评分矩阵的生成和评分预测，模型框架如图 1 所示。

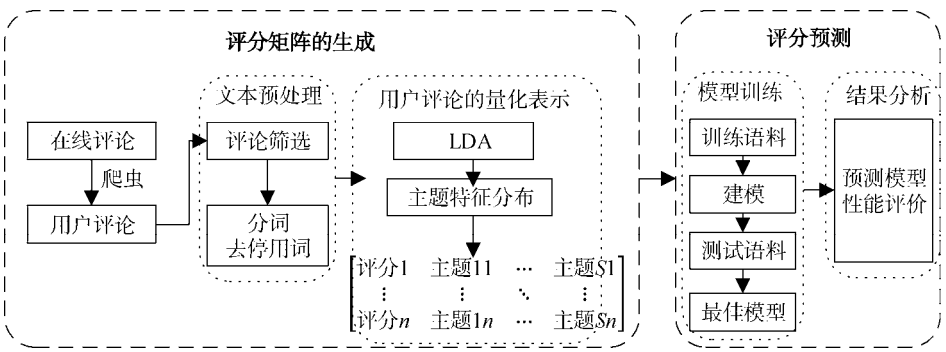

图 1　基于 XGBoost 算法的用户评分预测模型框架

3.1 用户评论量化

评分矩阵的生成是将用户的文本评论量化成结构化数据[15]。LDA[16]主题模型是由 Blei 等提出的包含词、主题和文档三层结构的贝叶斯概率模型，通过在文档和词之间引入主题维度，实现对向量空间的降维，可以处理大规模语料。该模型采用词袋的方法，将每篇文档视为一个词频向量，α 和 β 分别是文档的主题分布 θ 和主题中词的分布 Φ 的超参数，服从先验 Dirichlet 分布。本文使用 LDA 模型分析用户评论，识别评论中潜在的主题特征，估计各主题下词的概率分布。根据评论是否涉及各个主题，将用户评论量化成基于各主题特征的向量。

根据用户评论中出现的各主题词的数量及其概率分布，将用户评论量化成主题特征向量。记评论 u_i 量化后的主题特征向量为 θ_i，$\theta_i=[\theta_{i1},\cdots,\theta_{is}\cdots\theta_{iS}], s\in[1,S]$，其中，$S$ 是实验设置的主题个数，θ_{is} 表示评论 u_i 的第 s 个主题特征值。特征值的计算如公式(1)所示。

$$\theta_{is}=\sum_{v=1}^{V}(\varphi_{sv}\times n_{iv}) \tag{1}$$

其中，V 是各主题的主题词个数，φ_{sv} 表示第 s 个主题中主题词 v 出现的概率，n_{iv} 表示评论 u_i 中出现主题词 v 的次数，若评论中不包含主题词 v，则 n_{iv}=0。

3.2 XGBoost 算法集成

将包含 m 个特征、容量为 n 的数据集记为 $D=\{(x_i,y_i):x_i\in R^m, y_i\in R, |D|=n\}$，所有 CART 树的集合记为 $F=\{f(x)=w_{q(x)}, q:R^m\to T, w\in R^T\}$。其中，$q$ 代表样本映射到相应的叶子节点的决策规则，T 代表一棵树的叶子节点数量，w 代表叶子节点的得分。f 代表 CART 树，包括树的结构 q 和叶子节点的得分 w。基于 XGBoost 算法的 y_i 的预测值可以表示为公式(2)。

$$\hat{y}_i=\varphi(x_i)=\sum_{k=1}^{K}f_k(x_i) \tag{2}$$

其中，$f_k\in F$，K 为 CART 树的数量。XGBoost 算法在每一次模型训练时保留前面 $t-1$ 轮的预测不变，加入新函数 f_t 到模型中，$\hat{y}_i^t=\hat{y}_i^{(t-1)}+f_t(x_i)$ 为第 i 个样本在第 t 次模型训练时的预测结果。假设基学习器的误差相互独立，XGBoost 算法的学习目标是找到 f_t，最小化目标函数，其计算如公式(3)[12]和公式(4)[12]所示。

$$L^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)+\text{constant} \tag{3}$$

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{4}$$

其中，$l(\cdot,\cdot)$ 是训练误差，描述预测值与真实值之间差异的损失。$\Omega(\cdot)$ 是模型复杂度的正则项惩罚函数，γ 是复杂度参数，λ 是一个固定系数。

XGBoost 算法采用贪心算法[17]从根节点开始，递归地选择树结构的最优特征，据此特征对训练数据进行分割。假设 I_L 和 I_R 分别是分割点左边和右边的样本集，$I=I_L\cup I_R$。计算每个分割方案的信息增益，信息增益最大的分割为该节点的最优分割，其计算如公式(5)[12]所示。

$$L_{split}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{5}$$

其中，$I_j=\{i\mid q(x_i)=j\}$ 为节点 j 上的样本集合，$g_i=\partial_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$，$h_i=\partial^2_{\hat{y}_i^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$ 分别为训练误差的一阶和二阶梯度统计量。$\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}$、$\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}$、$\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}$ 分别为左子树分数、右子树分数、不分割时的分数，γ 为加入新叶子节点引入的复杂度代价，当 $L_{split}<0$ 时，放弃分割。

集成学习的关键是如何生成好而不同的个体学习器，个体学习器准确性越高、多样性越大，则集成效果越好。相对于 LR、NB 等稳定分类模型，树模型是一种对样本扰动比较敏感的不稳定分类模型。决策树因为其简单直观，具有很强的可解释性，常被作为集成学习的个体学习器。相比于 ID3 和 C4.5, CART[18]采用二元递归划分方法构建二叉树，分裂特征可以重复使用，既可以用于分类也可以用于回归。基于集成学习的优良性，本文在模型底层选用以 CART 树为基学习器的 XGBoost 算法建立评分预测模型。同时，为了增加集成学习中个体学习器的多样性，提高模型的泛化能力，在顶层利用 Bagging 思想加入数据样本扰动和属性扰动，以建立多个好而不同的 XGBoost 模型，并采用投票法进行模型集成，集成方法如图 2 所示。

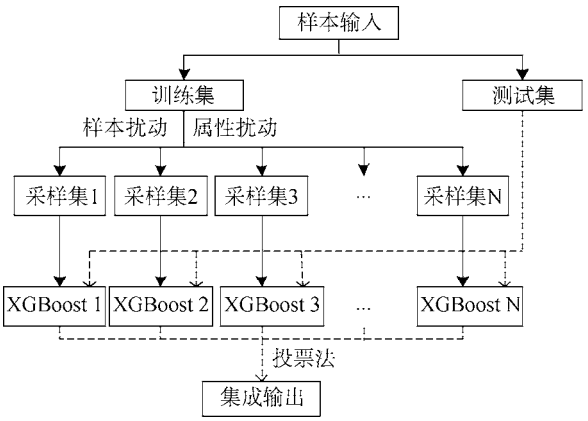

图 2　XGBoost 算法集成流程

对于二分类任务，类别标记集合记为$\{y_0, y_1\}$。将样本 x 在 XGBoost 模型 $\varphi_n(x)$ 上的预测输出表示为一个二维向量$\left\{\varphi_n^0(x), \varphi_n^1(x)\right\}$，则对样本 x，通过 N 个 XGBoost 模型集成进行预测，预测结果可以表示为公式(6)。

$$\hat{y}(x) = y_{\underset{j\in\{0,1\}}{argmax}\sum_{n=1}^{N}\varphi_n^j(x)} \tag{6}$$

3.3　评分预测模型构建

基于 XGBoost 算法的用户评分预测模型将用户评论的主题特征向量作为解释变量 x，用户评分作为被解释变量 y，具体预测步骤如下:

(1) 数据预处理

①收集用户在线评论，并进行筛选、分词、去停用词等数据预处理。

②采用 LDA 模型提取用户评论的主题特征及其概率分布，将用户评论量化成主题特征向量。

③以主题特征向量为解释变量，用户评分作为被解释变量构建样本集，并将样本集划分为训练集和测试集。

(2) 利用训练集建立基于 XGBoost 算法的用户评分预测模型

④加入样本扰动和属性扰动对训练集样本进行随机采样，生成多个采样集。

⑤对每一个采样集，重复步骤 1)–步骤 3):

1) XGBoost 算法根据公式(5)从根节点开始，递归地选择树结构的最优特征，据此特征对数据集进行划分，直到达到提前设定的划分停止条件(如树的最大深度等)，至此所有样本被分配到叶子节点，生成一棵 CART 树。

2) 重复步骤 1)，在损失函数梯度下降方向上依次建立多棵 CART 决策树。

3) 组合多棵 CART 决策树建立基于 XGBoost 算法的用户评分预测模型。

⑥对步骤⑤生成的多个评分预测模型利用投票法进行集成，生成最终的用户评分预测模型。

(3) 利用测试集对用户评分预测模型进行评价

⑦对于测试集的每条样本，利用步骤⑤生成的多个 XGBoost 模型，分别计算所有 CART 树叶子节点上的预测分数之和，若其属于正类的概率大于 0.5，将其划分为正类；否则将其划分为负类。

⑧根据公式(6)对多个 XGBoost 模型的预测结果进行集成，得到样本的预测类别。

⑨计算用户评分预测模型准确性指标：准确度(Accuracy)和 ROC(Receiver Operating Characteristic)曲线，评价评分预测模型的预测准确性。准确度是指分类正确的样本数占样本总数的比例。通常来说，分类准确度越高，算法越好。

4　用户评分预测

4.1　数据的收集和预处理

实验数据来自于国内某汽车门户网站的 2015 年 1 月至 2016 年 6 月的汽车用户在线评论，包括用户评分和评论数据。用户评分为两类: 1 为评分高, 0 为评分不高。利用汽车门户网站用户评论，预测用户评分，检验预测效果。

选取紧凑型 SUV 里评论热度较高的汽车品牌，这些品牌价格适中且购买量较大。首先抓取用户评论、评分及车型、价格等信息。从抓取到的数据中筛选出裸车价格在 15 万到 25 万之间的汽车用户评论数据，删除内容较少的评论。然后进行分词、去停用词、去标点符号等文本数据预处理操作。另外，在评论量化后的数据中，出现了少量在某些维度上为 0 的主题特征向量，这部分数据较少，对其进行了剔除处理。最终有效实验数据统计信息如表 1 所示，共计 13 628 条。表 1 的第 2-11 列为评论数最多的 10 种车型，共有 5 702 条评论(约 42%)。第 12 列为其他车型的 7 926 条评论(约 58%)。评分为 1 的用户占比约为 47%，评分为 0 的用户占比约为 53%。

表 1　紧凑型 SUV 汽车用户评论数据统计信息

评分	车型											总数
	1	2	3	4	5	6	7	8	9	10	其他	
0	656	523	396	330	256	235	233	147	180	212	4 110	7 278
1	556	600	157	152	174	187	172	237	169	130	3 816	6 350
合计	1 212	1 123	553	482	430	422	405	384	349	342	7 926	13 628

本文使用 LDA 模型对汽车用户评论进行主题分析，估计汽车用户评论的主题特征分布 θ，并根据评论是否涉及各个主题，将用户评论量化成一组主题特征向量。经实验验证，对于 LDA 模型的先验参数，分别设定为 α=0.2, β=0.1，迭代次数=100。当用户评论的主题数量 S=6 时主题划分较为清晰，6 个主题分别为：空间、动力、操控、油耗、外观、内饰。各主题中高频词汇云图如图 3 所示。

图 3　用户评论各主题中高频词汇云图

根据图 3 实验结果，用户关注的紧凑型 SUV 车型特征为：外观方面关注是否大气、时尚；动力方面关注起步、油门、加速等；空间方面关注后排和后备箱等空间是否够用；内饰方面关注做工、座椅、用料等；油耗方面关注跑市区和高速时的百公里油耗；操控方面关注方向盘是否轻盈、转向是否精准、底盘是否稳健。

选取每个主题中出现概率较高的前 30 个代表性词汇，即 V=30。根据评论是否涉及各个主题词，将用户评论量化成主题特征向量，作为解释变量 x，被解释变量为用户评分 y。量化后的汽车用户评论如表 2 所示。

表 2　用户评论的主题特征向量

评分	空间	动力	操控	油耗	外观	内饰
1	0.739 4	0.177 5	0.286 8	0.333 3	0.165 8	0.265 4
1	0.010 6	0.019 7	0.010 5	0.036 9	0.008 3	0.006 1
0	0.030 3	0.005 0	0.012 1	0.081 5	0.066 2	0.023 2
1	0.014 7	0.023 7	0.038 5	0.120 3	0.015 0	0.009 2
0	0.012 5	0.043 6	0.032 7	0.024 5	0.026 1	0.010 0
0	0.130 1	0.014 3	0.013 4	0.057 6	0.090 9	0.003 8
1	0.026 4	0.015 4	0.025 9	0.014 5	0.026 0	0.099 6
1	0.020 5	0.034 7	0.041 4	0.011 0	0.031 8	0.010 0
0	0.010 6	0.008 4	0.031 4	0.053 1	0.151 7	0.006 2
…	…	…	…	…	…	…

4.2　基于 XGBoost 算法的用户评分预测模型构建

从量化后的 13 628 条有效数据中，随机抽取 70% 作为训练集，30%作为测试集。训练集和测试集中 y=1 的数据比例相等。XGBoost 算法在训练过程中，调优参数扰动范围为：学习率 0.1~0.3，树的最大深度 5~10，样本采样比 0.7~1，样本属性采样比 0.7~1，迭代次数 100~1 000，正则化项权重为 5~10。XGBoost 算法的可解释性主要体现在特征重要性排序和决策树图两个方面。

选择用户评论的 6 个主题特征，同时考虑到主题特征之间可能会存在交互效应，即一个主题特征在另一个主题特征不同水平上可能会产生不同效果，实验设计中增加了 6 个主题特征的二次乘积项，用以表达不同主题特征之间的交互效应。用 Logistic 回归方法筛选出对模型贡献较高的 22 个特征，选择在生成所有树的过程中特征被用作分裂特征的次数作为特征权重，特征重要性排序如图 4 所示。特征重要性排序可以理解为大多数用户对紧凑型 SUV 不同角度的偏好，由图4可以看出，大多数用户对价格在15-20万之间的紧凑型 SUV 比较注重外观、油耗、动力、内饰等因素，用户在选车时考虑外观的同时会考虑空间、操控，考虑动力的同时也会考虑内饰等其他交互因素。

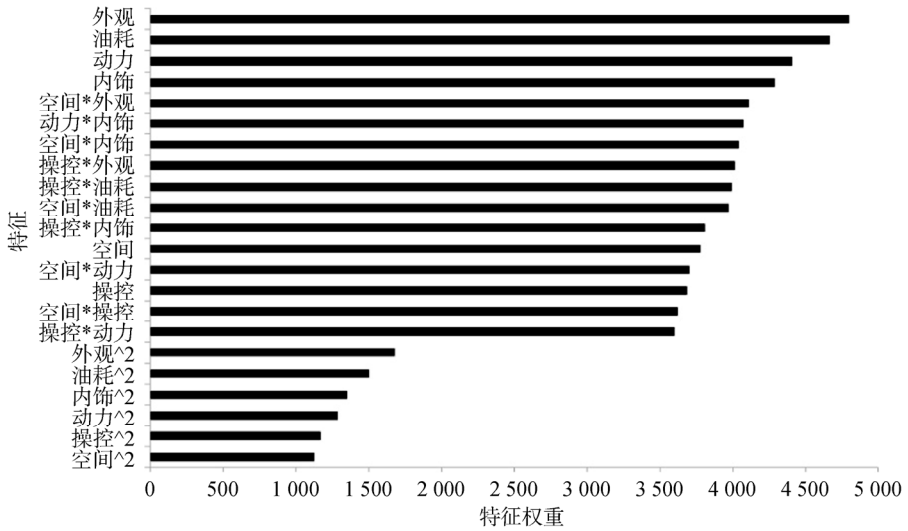

图 4　特征重要性排序

经实验发现，迭代次数为 1 000 时效果最优，1 000 棵 CART 树作为最终的用户评分预测模型。对于测试集的每条样本，利用生成的 1 000 棵 CART 树的决策规则对其进行预测，计算所有 CART 树对应叶子节点上的预测分数之和，若其属于正类的概率大于 0.5，预测评分为 1；否则预测评分为 0。XGBoost 算法生成的第 1000 棵 CART 树如图 5 所示。

为了演示用户评分预测过程，选取测试集中一条汽车用户评论。该用户对这款车的评分类别为 1。量化后的主题特征向量见表 2 中第一条样本。

外观大气漂亮，内饰做工精细，空间大，油耗低。后排乘坐非常舒适。车内空间宽敞，座椅的包裹性不错，后排座椅空间还是蛮大，乘坐舒适，后备箱空间有点小，美中不足的是全尺寸备胎导致的后箱隆起。动力不错，2.0 开起来一点也不弱，无论是在市里还是在高速上，……。

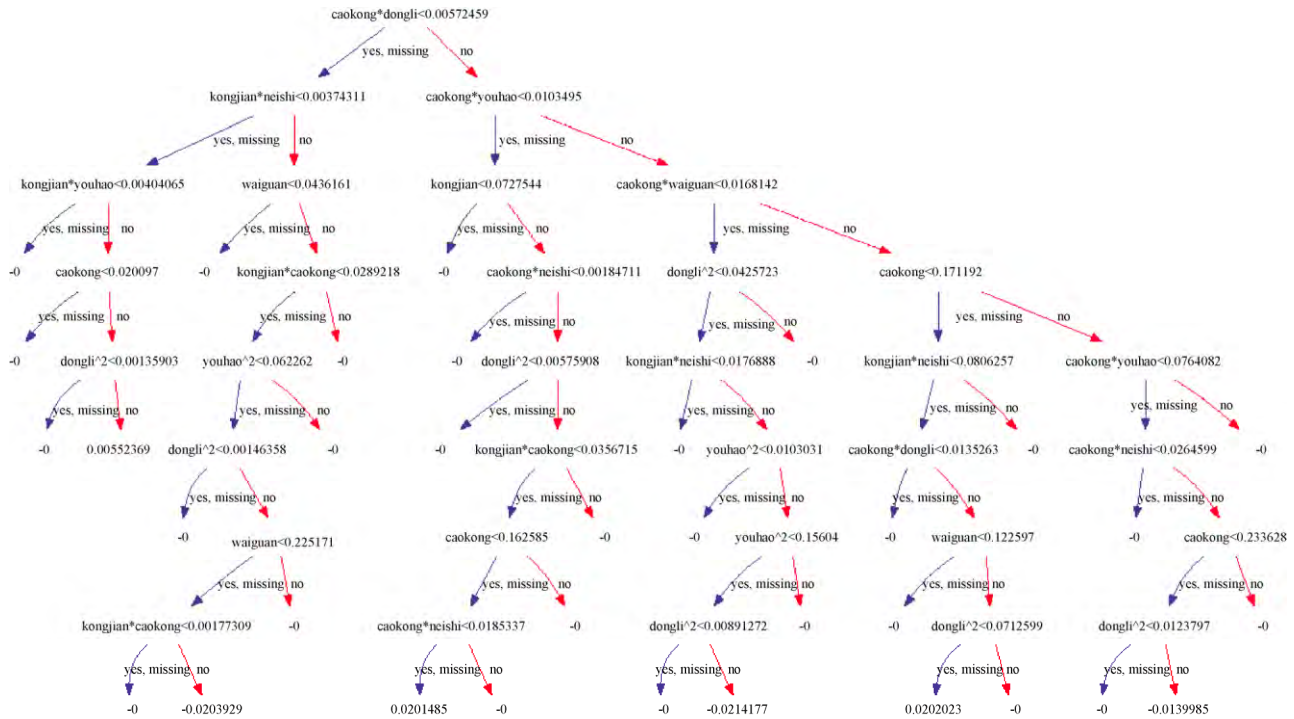

图 5　基于汽车用户评论数据的 XGBoost 决策图(第 1000 棵 CART 树)

依据图 5 第 1 000 棵 CART 树的决策规则，该样本被划分到最右侧叶子节点，得分为 0。同样利用第 1-999 棵 CART 树对该样本进行预测，预测得分如图 6 所示。

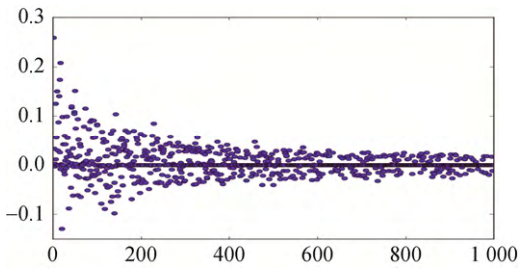

图 6　1000 棵 CART 树得分的散点图

该条样本得分的最小值低于–0.1，最大值高于 0.2，波动幅度较大。随着 CART 树数量的增加，波动幅度逐渐较小，趋于 0 值附近。总体来看，得分为正值的情况相对更多。将 1 000 棵 CART 树的叶子节点得分相加为 4.650 830 244，用 Logistic 函数进行转换，得到其类别为 1 的概率为 0.990 536 742，概率值大于 0.5，其预测评分类别为 1。对比图 3 可以看出，该评论中出现了外观、漂亮、大气、座椅、内饰、空间、后排等高频词汇。用户评分预测类别与真实类别一致。

4.3　评分预测模型准确性的评价

本文选用准确度、ROC 曲线对基于 XGBoost 算法的用户评分预测模型准确性进行评价。为了验证本文方法的优良性，分别选择对数据扰动不敏感的稳定分类算法 NB、LR，以及对数据扰动敏感的树类模型 CART、RF 算法进行比较。各分类算法预测结果的 ROC 曲线如图 7 所示。纵轴是 TPR，横轴是 FPR。

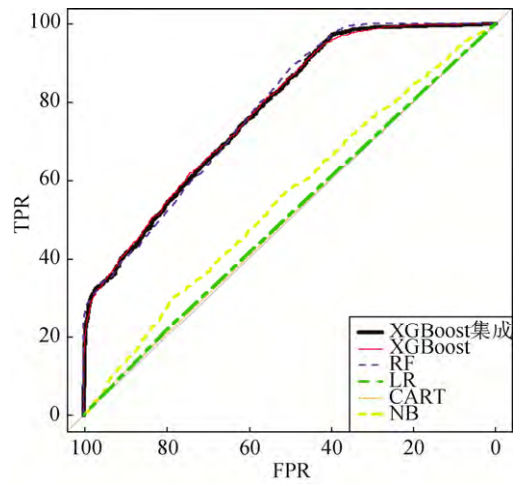

图 7　ROC 曲线比较

XGBoost 集成、XGBoost 和 RF 算法的 ROC 曲线明显包住 NB、LR、CART 树，说明 XGBoost 集成、XGBoost 和 RF 的预测准确性明显优于 NB、LR、CART 算法。

为了进一步比较预测准确度，对量化后的 13 628 条有效数据，采用 5 次五折交叉验证，分别建立评分预测模型，结果如表 3 所示。第 2 列至第 6 列分别为五折交叉验证的预测准确度，第 7 列为平均值。表 3 显示，XGBoost 集成、XGBoost 和 RF 算法的预测准确度明显优于 NB、LR、CART 算法。由平均值可知，XGBoost 算法的预测准确度达到 67.75%，比 RF 算法的预测准确度 67.54%高出约 0.21 个百分点。通过加入样本扰动和属性扰动对 XGBoost 算法集成以后，模型的泛化能力有所提高，预测准确度在 XGBoost 基础上提高约 0.43%，比随机森林高约 0.64%。表 3 中各方法的预测准确度与图 7 结论一致。

表 3　五折交叉验证的预测准确度比较

算法	1	2	3	4	5	平均
XGBoost 集成	69.31%	68.21%	66.86%	68.06%	68.47%	68.18%
XGBoost	69.15%	68.06%	65.96%	67.45%	68.14%	67.75%
RF	68.97%	68.10%	65.63%	66.86%	68.16%	67.54%
CART	54.75%	54.27%	52.51%	53.15%	55.17%	53.97%
LR	55.34%	54.53%	52.11%	54.84%	55.43%	54.45%
NB	53.50%	54.42%	50.50%	53.04%	54.07%	53.11%

由于样本数据规模略大，计算速度会受算法影响。通过实验证明，XGBoost 算法在迭代次数为 1 000 时，预测准确度趋于最大值，建模过程平均耗时约 9 秒。对 XGBoost 算法集成以后，迭代次数为 500 时，预测准确度趋于最大值，建模过程耗时约 50 秒。随机森林算法的决策树数量为 500 时，预测准确度趋于最大值，建模过程平均耗时约 120 秒。XGBoost 算法及其集成的计算速度明显优于随机森林。综上可知，基于 XGBoost 算法集成的用户评分预测模型具有较高的预测准确度和较高的计算效率。

4.4　小　结

选取国内某汽车门户网站的汽车用户在线评论数据，进行数据清洗和用户评论的主题特征提取及量化，构建基于 XGBoost 算法的用户评分预测模型。数据分析结果显示，该模型能够将用户评论信息与评分有效结合，揭示出用户对紧凑型 SUV 汽车的消费偏好，其

预测准确度高于 LR、NB 等模型，且计算效率的优势更为明显，能够更好地了解汽车用户需求，为汽车行业整体结构的优化提供参考。

5　总结与展望

本文将用户评论与评分相结合，利用LDA模型挖掘用户评论中的主题特征，并将评论量化成基于各主题特征的向量作为解释变量，将用户评分作为被解释变量，并在传统 XGBoost 算法的基础上加入数据样本扰动和属性扰动以建立多个不同的 XGBoost 模型，通过投票法对模型进行集成，提出一种基于 XGBoost 算法的用户评分预测模型。以国内某汽车门户网站用户口碑评论数据为研究对象，对汽车产品的真实用户评论数据进行主题特征提取和建模，并与 NB、LR、RF 等算法进行对比，实验结果表明基于 XGBoost 集成的用户评分预测模型具有较好的预测准确度和计算效率，泛化能力较强，适用于处理大规模数据。

在未来研究中，将分别从用户与商品角度挖掘用户偏好和建立商品画像，用于商品推荐的评分预测。此外，还将继续寻找 XGBoost 算法集成的参数优化和模型融合方法，以进一步提高 XGBoost 算法集成的分类准确性和计算效率，完善用户评分预测模型及其应用。

参考文献:

[1] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems[J]. Computer, 2009, 42(8): 30-37.

[2] Koren Y, Bell R. Advances in Collaborative Filtering[A]// Recommender Systems Handbook[M]. New York: Springer, 2011: 145-186.

[3] 邓晓懿，金淳，韩庆平，等. 基于情境聚类和用户评级的协同过滤推荐模型[J]. 系统工程理论与实践, 2013, 33(11): 2945-2953. (Deng Xiaoyi, Jin Chun, Han Jim C, et al. Improved Collaborative Filtering Model Based on Context Clustering and User Ranking[J]. Systems Engineering—Theory & Practice, 2013, 33(11): 2945-2953.)

[4] Li X, Xu G, Chen E, et al. Learning User Preferences across Multiple Aspects for Merchant Recommendation[C]// Proceedings of the 2015 IEEE International Conference on Data Mining. IEEE, 2015.

[5] Fan M, Khademi M. Predicting a Business Star in Yelp from Its Reviews Text Alone[OL]. arXiv Preprint, arXiv: 1401.0864.

[6] 张红丽，刘济郢，杨斯楠，等. 基于网络用户评论的评分预测模型研究[J]. 数据分析与知识发现, 2017, 1(8): 48-58. (Zhang Hongli, Liu Jiying, Yang Sinan, et al. Predicting Online Users' Ratings with Comments[J]. Data Analysis and Knowledge Discovery, 2017, 1(8): 48-58.)

[7] 高祎璠，余文喆，晁平复，等. 基于评论分析的评分预测与推荐[J]. 华东师范大学学报：自然科学版，2015(3): 80-90. (Gao Yifan, Yu Wenzhe, Chao Pingfu, et al. Analyzing Reviews for Rating Prediction and Item Recommendation[J]. Journal of East China Normal University: Natural Science, 2015(3): 80-90.)

[8] 杨博，赵鹏飞. 推荐算法综述[J]. 山西大学学报：自然科学版, 2011, 34(3): 337-350. (Yang Bo, Zhao Pengfei. Review of the Art of Recommendation Algorithms[J]. Journal of Shanxi University: Natural Science Edition, 2011, 34(3): 337-350.)

[9] Brown I, Mues C. An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets[J]. Expert Systems with Applications, 2012, 39(3): 3446-3453.

[10] 应维云. 随机森林方法及其在客户流失预测中的应用研究[J]. 管理评论，2012, 24(2): 140-145. (Ying Weiyun. The Research on Random Forests and the Application in Customer Churn Prediction[J]. Management Review, 2012, 24(2): 140-145.)

[11] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[13] Seyfioğlu M, Demirezen M. A Hierarchical Approach for Sentiment Analysis and Categorization of Turkish Written Customer Relationship Management Data[C]//Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. IEEE, 2017: 361-365.

[14] Athanasiou V, Maragoudakis M. A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources are Not Plentiful: A Case Study for Modern Greek[J]. Algorithms, 2017, 10(1): 34.

[15] Zhang R, Gao Y, Yu W, et al. Review Comment Analysis for Predicting Ratings[A]// Web-Age Information Management [M]. Springer, 2015: 247-259.

[16] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J].

Journal of Machine Learning Research, 2003, 3: 993-1022.

[17] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[18] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees (CART)[J]. Encyclopedia of Ecology, 1984, 40(3): 582-588.

作者贡献声明:

杨贵军, 徐雪: 提出研究思路, 设计研究方案, 论文最终版本修订;
徐雪: 进行实验, 论文起草;
赵富强, 徐雪: 采集、清洗和分析数据。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: xuxue2017@163.com。

[1] 杨贵军, 徐雪, 赵富强. ASUVp(15-25)_comment.csv. 汽车网站用户评论数据.

[2] 杨贵军, 徐雪, 赵富强. LDA_twords.dat. LDA 实验结果.

[3] 杨贵军, 徐雪, 赵富强. train.csv, test.csv. 训练集与测试集数据.

[4] 杨贵军, 徐雪, 赵富强. fold1.csv, fold2.csv, fold3.csv, fold4.csv, fold5.csv. 五折交叉验证数据.

[5] 杨贵军, 徐雪, 赵富强. LDA 源代码. docx. Python 下 LDA 源代码.

[6] 杨贵军, 徐雪, 赵富强. XGBoost 源代码. docx. Python 下 XGBoost 源代码.

收稿日期: 2018-04-13
收修改稿日期: 2018-07-02

Predicting User Ratings with XGBoost Algorithm

Yang Guijun[1] Xu Xue[1] Zhao Fuqiang[2]
[1](China Center of Economics and Statistics Research, Tianjin University of Finance and Economics, Tianjin 300222, China)
[2](Institute of Polytechnic, Tianjin University of Finance and Economics, Tianjin 300222, China)

Abstract: [Objective] This study aims to build a model for effectively predicting ratings of user reviews and analysing consumer behaviours. **[Methods]** First, we applied the Latent Dirichlet Allocation model to set the topic features from user reviews as independent variable and user ratings as dependent variable. Then, we built a user rating prediction model based on the eXtreme Gradient Boosting algorithm. Finally, we added the disturbances of samples and attributes to the proposed model for rating prediction. **[Results]** We used the new model to predict user's comments on a domestic automobile online portal, and identified their preferences of automobile. Compared with the Logical Regression and Random Forest algorithms, the proposed model has better precision and efficiency. **[Limitations]** We need to include data from other fields to more comprehensively describe user's behaviours. **[Conclusions]** The proposed model could quantify user's reviews and then predict their ratings effectively.

Keywords: Rating Prediction XGBoost Algorithm LDA Feature Extraction User Reviews