一种融合历史均值与提升树的客流量预测模型

白智远,温从威,杨锦浩,陈 智,吕 品
(上海电机学院 电子信息学院,上海 201306)

摘 要:移动定位服务的发展使得互联网商家“线上线下”的交易数据急剧增长,如何挖掘出海量交易数据中隐藏的用户行为、实现智能化决策是互联网商家在运营过程中面临的一个重要问题。基于此,提出了一种融合历史均值与提升树的互联网商家客流量预测模型,其中提升树用于改进模型的预测精度,历史均值模型用于考虑客流量预测与时间的依赖关系。历史均值与提升树融合的核心思想是先通过提升树 XGBoost、GBDT 和历史均值模型预测商家过去三周的平均销量和总销量,然后,构建提升树模型与历史均值模型的融合权重系数公式。在包含 2 000 个互联网商家销售数据集上实现了该方法,并将其与时间序列加权回归模型进行了对比,发现两种方法的预测结果相似,这表明该方法考虑时间因素是正确合理的;并且在训练集增大的情况下,模型的预测精度能得到显著改善。

关键词:历史均值;提升树;时间序列加权回归;互联网商家;客流量
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2019)04-0212-04
doi:10.3969/j.issn.1673-629X.2019.04.043

A Passenger Flow Predication Model Combining History Means and Boosting Tree

BAI Zhi-yuan, WEN Cong-wei, YANG Jin-hao, CHEN Zhi, LYU Pin
(School of Electronics and Information, Shanghai Dianji University, Shanghai 201306, China)

Abstract: The development of mobile positioning service makes the online and offline transaction data of Internet merchants grow rapidly. How to dig out the hidden user behaviors in the massive transaction data and realize the intelligent decision-making is a critical problem that Internet merchants are facing in the process of operation. Based on this, we propose an Internet merchant traffic prediction model integrating historical mean and boosting tree, in which the boosting tree is used to improve the prediction accuracy, and the historical mean model is used to consider the dependence between passenger flow prediction and time. The core idea of the proposed model is to predict the average sales and total sales of merchants in the past three weeks by XGBoost, GBDT and historical mean model, and then build the fusion weight coefficient formula of the boosting tree and historical mean model. This method is implemented on the sales data set of 2 000 Internet merchants, and compared with the weighted regression model of time series. It is found that the results of the two methods are similar, which indicates that the proposed method is correct to consider the time factor. Moreover, with the increase of training set, the prediction accuracy of the model can be significantly improved.

Key words: history mean; boosting tree; time series weighting regression; Internet business; passenger flow

0 引 言

移动定位服务的发展使得互联网商家“线上线下”的交易数据急剧增长[1-4]。分析这些数据中隐藏的用户交易习惯和倾向性[5-6]对优化商家的运营具有重要作用。近来年,出现了许多关于移动定位服务预测的研究。例如,付全兴等[7]使用逻辑回归和支持向量机,以 4 个月的电商数据为研究对象,预测用户的购买行为;陈传波等[8]把平滑加权的思想应用于实时模型预测,通过提取包含有趋势的特征来提高预测模型的精确度;张昊等[9]利用 XGBoost(extreme gradient boost)算法[10]实现了商品推荐中的用户购买行为预测。他们将决策树[11]、随机森林[12]作为基线对比方法,研究发现变量的重要性对模型的构建有较大影响。

文中借鉴上述研究的思想,提出了历史均值与提升树融合的互联网商家客流量预测模型。该模型的本质是提升树模型与历史均值模型,按照计算公式所求

收稿日期:2018-02-07 **修回日期**:2018-06-12 **网络出版时间**:2018-12-20
基金项目:2017 年上海市大学生科创项目(A1-5701-17-009-02-54);上海市教育科学研究项目(C17014/17AR04)
作者简介:白智远(1996-),男,研究方向为数据挖掘与机器学习;吕 品,副教授,博士,CCF 会员(60050M),研究方向为数据挖掘、情感分析。

出的权重系数,按照一定比例而融合的加权和,不仅考虑了如何提高模型的预测精度,而且还考虑了客流量的预测与时间的依赖关系。并且对不同模型的预测结果做出了对比分析。最后,将融合了历史均值与提升树的客流量预测模型所得到的结果与传统的零售业结合,粗略进行了分析,对商家今后的运营提出了一些实质性的建议。

1　数据预处理

1.1　数据描述

文中使用的数据来自天池大数据平台,共包含某年 7 月 1 日至次年 10 月 31 日的商家完整行为数据,分为“商家特征”数据、“用户支付行为”数据和“用户浏览行为”数据。商家特征反映了商家的热度,评分高以及评论好的商家,是提高用户购买力的因素之一,除此之外,门店的等级、菜品的丰富程度也作为商家的考量之一。它的数据共包含 7 个属性:商家 ID、店铺所在地、人均消费、评分、评论数、门店等级以及食品分类名称;用户支付行为特征反映了用户的支付习惯方式,包含 3 个属性:用户 ID、商家 ID 和用户的支付时间;用户浏览行为则反映了用户的购买习惯,如果用户经常访问同一个商家,结合其他两个特征可以推断出用户所喜爱的商品种类、个人口味等信息,包含 3 个属性:用户 ID、商家 ID 和用户浏览商家的时间。

1.2　数据预处理方法

由于直接使用原始数据训练模型不仅会产生误差,还会耗费大量的计算资源,因此,对原始数据集进行了预处理,将原始数据中存在的异常值进行剔除、去重、归一化等处理。一方面,由于商家从入驻口碑平台到销售量增加存在一定的启动时间,并且可能出现某段时间销量中断的现象,因此,商家开业前 7 天的数据以及销量中断前后 3 天的数据不作为训练数据;另一方面,由于原始数据中存在短时间内单个用户大量购买的情况,为消除这种异常消费对预测的影响,采用了基于规则的方法对原始数据进行归一化;另外,原始数据中还存在一些特殊时间节点和难以预计的大幅波动,如大型节假日(如中秋节、国庆节等)、停业、商家开展促销活动时单个用户大量购买的情况。对于这些基于规则的方法难以处理的异常值,文中采用了模型预训练方法,即采用欠拟合算法对模型预训练,清除原始数据中残差为 10% 和 25% 的数据。由于预测目标是商家的日销量,因此预处理后用于训练的数据是按小时统计的商家的总销量。

此外,为提高模型预测的准确性,实验中还采集了全国各省市的天气数据以及节假日天气数据作为原始数据的补充。在额外采集的气温、湿度、气压等数据中,根据经验,将天气状况简单转换为降水指数和天晴指数两个指标。由于人体对于气象参数的感受不成线性关系,故生成人体舒适度指数(comfort index of human body,SSD)作为模型训练的一个重要特征。最终,模型训练与预测使用的特征与标签如表 1 所示。

表 1　模型训练与预测使用的特征

特征	描述
历史销量	过去 21 天的历史销量
节假日	过去 21 天及预测 14 天的节假日标注
销量	工作日标注为 0,周末标注为 1,假期标注为 2
天气	过去 21 天及预测当天附近 4 天(之前两天,当天,之后一天)的降水量,人体舒适度 SSD 值,降水指数,天晴指数
商家	平均 View/Pay 比值,平均每天开店时间,关店时间,开店总时长;首次营业日期,非节假日销量中位数,节假日销量中位数,节假日/非节假日销量比值;商家类别,人均消费,评分,评论数,门店等级

2　历史均值与提升树融合的客流量预测

2.1　XGBoost 的基本思想

XGBoost 是一种极限提升树的机器学习方法,具有良好的扩展性,以及计算速度快、模型表现好等特点。对于数据集 $D=\{(x_i,y_i)\}$,提升树方法的核心是最小化式 1 所示的正则化目标函数。

$$L(\varphi)=\sum_i l(\hat{y}_i,y_i)+\sum_k \Omega(f_k) \tag{1}$$

其中,$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2$,$l$ 是可微的凸损失函数,用来衡量预测值 $\hat{y}_i$ 和目标 y_i 之间的差异;$\Omega(f_k)$ 表示模型的复杂度。

一般,对上述目标函数进行二阶泰勒展开(如式 2),然后进行优化。

$$L^{(t)}\simeq\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{(t-1)})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{2}$$

其中,$g_i=\partial_{\hat{y}^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$,$h_i=\partial^2_{\hat{y}^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$。

假设树结构 $q(x)$ 已知,并且 $I_j=\{i\mid q(x_i)=j\}$ 为叶节点 j 的样本集合,可得叶节点 j 的最优权重:

$$w_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \tag{3}$$

最后,采用贪心算法,从某一叶子开始,反复向树中添加分支。假设 I_L 和 I_R 是分割后左右节点的实例集合。令 $I=I_L\cup I_R$,则分裂后的损失可由式 4 计算。

$$L_{\text{split}}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}\right]-\gamma \tag{4}$$

与传统的 GBDT 模型对比,XGBoost 还支持线性分类器,并且加入正则化因子,用于控制模型的复杂度。正则项里包含了树的叶子节点个数等信息,它降低了模型的方差,使学习出来的模型更加简单,防止过拟合,这也是 XGBoost 优于传统 GBDT 的一个特性。

2.2 历史均值模型的基本思想

历史均值模型是以预测日为基准,求出预测日之前到某一天的平均客流量、销量增量等信息,再以权重系数作为融合的比例,预测未来 14 天的客流量。

2.3 融合方法

为获得精确度高的客流量预测模型,文中采用了二个阶段的训练方法。第一次阶段的训练中,使用了 XGBoost 与 GBDT(gradient boosting decision tree)模型。模型训练的参数如表 2 和表 3 所示。每一种模型分别使用了 2 组参数进行训练,总共获得 4 个模型。

表 2 XGBoost 算法的不同参数

XGBoost	1 号	2 号
目标函数	线性回归模型	线性回归模型
树的最大深度	3	5
学习率	0.1	0.03
提升树个数	500	1 600
L1 正则化项参数	0	1
L2 正则化项参数	1	0

表 3 GBDT 算法的不同参数

GBDT	树的最大深度	学习率	提升树个数	训练采样比例
1 号	3	0.1	500	0.95
2 号	5	0.1	500	0.95

为了减小预测误差,调整 XGBoost 与 GBDT 算法中树的深度、学习率以及迭代次数的参数,在 XGBoost 算法的 1 号模型中,一般情况下,学习率的值默认为 0.1,而树的最大深度默认为 3。但是,对于不同的问题,理想的学习率有时会在一些特定的区间范围之间波动。树的深度越大,则对数据的拟合程度越高。因此,文中在确定 XGBoost 算法的 2 号模型的学习率以及树的最大深度时,引入 XGBoost 算法中内置的 cv 函数,cv 函数在每一轮迭代中使用交叉验证,根据算法参数的调整,返回理想的决策树数量。因此,通过 cv 函数较为精确的计算,将 2 号模型的学习率调至 0.03,树的最大深度为 5。

第二阶段的训练使用了历史均值模型。历史均值模型以预测日为基准,首先求出预测日之前的 21 天的销量平均值,得到每天的平均销量;其次,以周为单位,统计每周的销量的中位数和平均值,通过线性拟合得到每周的销量增量;最后,将每天的均值销量与每周的销量增量叠加,以此预测未来两周的销量。该模型把过去 21 天的历史销量的相关度矩阵作为输入;将未来两周的销量和历史均值模型与第一阶段的模型融合的权重系数作为输出。均值模型的融合比例最大为 0.75。融合的权重系数计算如下:

$$\text{Credit} = 0.75 \times \frac{\min(\overline{\text{Fus}}, \text{Fus}_{\text{last}}) - 0.7}{0.3} \tag{5}$$

其中,$\overline{\text{Fus}}$ 是过去三周的平均销量;Fus_{last} 为过去三周的销量。

由此,将 XGBoost、GBDT 和历史均值模型得到的过去三周的平均销量和销量值,分别代入式 5,可求出相应的权重系数为:0.47,0.34,0.19。最终,将训练得到的 2 组 XGBoost 模型和 2 组 GBDT 的不同结果分别与历史均值模型按 0.47,0.34,0.19 的比例融合,得到预测未来 14 天的客流量。

3 实验分析

3.1 实验设置

该实验采用的硬件为 Inter(R) Core(TM) i5-5200U CPU @ 2.20 GHz。软件环境中操作系统为 Windows 7,开发环境为 Python3.6。原始数据为 2.13 GB,预处理后的数据为 220 MB。为判断 XGBoost 方法预测的有效性,实验中引入了时间序列加权回归的算法作为基线对比方法[8]。

3.2 预测结果对比分析

由于时间序列反映了实体属性在时间顺序上的特征[13],因此,实现了时间序列加权回归算法,分析 2 种算法的预测结果后,得到的前 500 位互联网商家在未来 14 天的客流量发展趋势,如图 1 和图 2 所示。

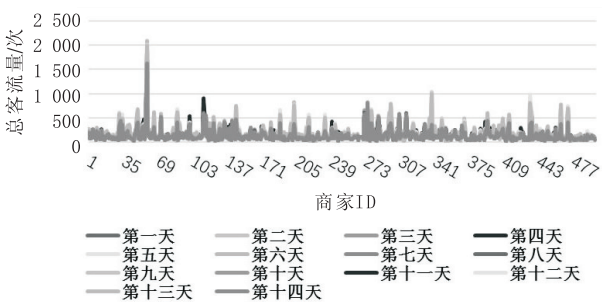

图 1 历史均值与提升树融合模型预测

分析客流量发展趋势可知:

(1)与浏览动作相关的变量对模型的贡献程度最大,这是因为浏览是用户交互的最主要方式,其信息丰富程度远高于其他特征;

(2)部分商家可能所经营的商品评价较高,顾客的返回率使得部分商家的客流量稳步上升;

(3)大部分的商家十四天总客流量已经突破了 5 000,少量甚至达到了约 25 000 的级别。这极有可能是商家近期的某种促销活动所导致的。比如通过平台

派发不同程度的优惠券、现金红包、买满一定金额优惠等活动。但如何调整自己的运营策略,吸引到更多的客流量显得至关重要。

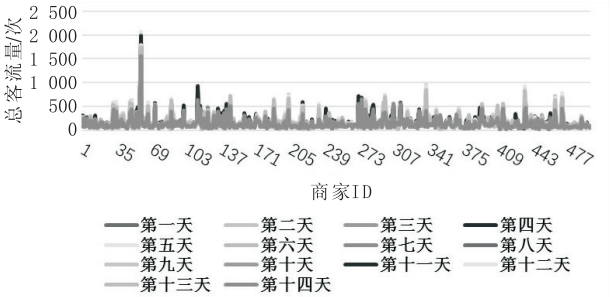

图 2　时间序列加权回归模型预测

3.3　性能分析

通过优化算法参数,采用测试集样本对建模结果进行评测[14],算法运行结果和精度测试如表 4 所示。

表 4　历史均值与提升树融合模型精度测试

训练集样本大小	运算时间/s	fevalerror
500	213.101	0.043 16
1 000	314.215	0.040 15
1 500	381.219	0.038 45
2 000	416.125	0.036 86

实验中利用 XGBoost 自定义的评价函数对提出的模型进行了性能评估。调用评价函数时,传入验证集和验证集上的预测值作为函数参数,返回一个浮点类型的评估值 fevalerror。fevalerror 的值越大,模型预测精度越低。反之,fevalerror 的值越小,模型预测精度越高。结果表明,随着训练集样本大小的增加,运算时间增加,fevalerror 值逐渐减小,精度上却逐渐增加。由此,历史均值与提升树的融合模型具有预测精度较高、运算速度较快的优势。

4　结束语

将历史均值模型与提升树方法进行了融合,对互联网商家的线上线下的真实用户数据进行了特征提取和建模预测。并将提出的模型与时间序列加权回归进行了预测结果与性能比较。实验结果表明,融合历史均值模型与提升树模型的方法能有效实现互联网商家客流量的预测。在互联网高速发展的今天,对比传统的零售行业,互联网商家的营销对用户消费给予了更多的关注,在产品详情页的介绍、客服服务、便捷的移动支付等方面都致力于为用户带来更好的消费体验。通过这次客流量预测模型的构建和对用户数据进行的挖掘,商家利用互联网这一渠道,能够更好地与用户及时沟通,了解用户感受,使互联网商家与用户建立了信任关系,吸引到更多忠实的用户。这对互联网商家的运营决策、降低成本、改善用户体验有着重要的现实意义。

参考文献:

[1] 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.

[2] KIM J, HAN Jiawei, YUAN Cangzhou. TOPTRAC: topical trajectory pattern mining[C]//Proceeding of 2015 ACM SIGKDD international conference on knowledge discovery and data mining. Sydney, NSW, Australia: ACM, 2015: 587-596.

[3] ZHANG Chao, ZHOU Guangyu, YUAN Quan, et al. GeoBurst: real-time local event detection in geo-tagged tweet stream[C]//Proceeding of 2016 ACM SIGIR conference on research & development in information retrieval. Pisa, Italy: ACM, 2016: 513-522.

[4] ZHANG A, GOYAL A, BAEZA-YATES R, et al. Towards mobile query auto-completion: an efficient mobile application-aware approach[C]//Proceeding of 25th international conference on world wide web. Montréal, Québec, Canada: WWW, 2016: 579-590.

[5] GUI Huan, LIU Haishan, MENG Xiangrui, et al. Downside management in recommender systems[C]//Proceeding of 2016 IEEE/ACM international conference on advances in social networks analysis and mining. San Francisco, CA, USA: IEEE, 2016: 394-401.

[6] QU Meng, TANG Jian, SHANG Jingbo, et al. An attention-based collaboration framework for multi-view network representation learning[C]//Proceeding of 2017 ACM international conference on information and knowledge management. Singapore: ACM, 2017: 1767-1776.

[7] 付全兴,韩立新,杨　艺. 基于生活场景的逻辑回归推荐算法[J]. 计算机与现代化,2016(12):38-41.

[8] 陈传波,潘　非,李其申,等. 时间序列趋势加权平滑预测模型研究[J]. 小型微型计算机系统,2001,22(11):1299-1301.

[9] 张　昊,纪宏超,张红宇. XGBoost 算法在电子商务商品推荐中的应用[J]. 物联网技术,2017,7(2):102-104.

[10] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco: ACM, 2016: 785-794.

[11] 曹颖超. 改进的 GDBT 迭代决策树分类算法及其应用[J]. 科技视界,2017(12):105.

[12] 王爱平,万国伟,程志全,等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报,2011,22(9):2059-2074.

[13] 肖　瑞,刘国华. 基于趋势的时间序列相似性度量和聚类研究[J]. 计算机应用研究,2014,31(9):2600-2605.

[14] 殷春霞,楚　涛,马　力. 基于数据挖掘的网络性能分析系统的设计和实现[J]. 计算机工程,2006,32(12):136-138.