

线性回归和随机森林算法融合在餐饮客流量的预测

杨森彬

(广东工业大学自动化学院, 广东 广州 525000)

摘要:数据挖掘技术运用于餐饮行业具有一定的社会价值,通过预测餐饮行业客流量,根据客流量多少餐厅合理为顾客准备用餐,有利于提升顾客用餐体验,提高餐饮质量的同时让餐饮行业更高效运作。本文通过研究线性回归算法与随机森林算法理论,提出将线性回归算法与随机森林算法融合的思想,将其应用在餐厅顾客回访数量预测,并通过实验证明该思路的合理性和可实施性。通过实验对比,算法融合思路比线性回归算法准确率提高了约3.004%,比随机森林算法提高了约2.022%。比以往大部分研究取得更优的预测效果,为数据挖掘技术在餐饮行业的应用提供了新的思路。

关键词:数据挖掘;线性回归;随机森林;算法融合;餐饮行业

中图分类号: TP312 **文献标识码:** A

Combining Linear Regression Algorithm with Random Forest Algorithm to Predict Numbers of Restaurant Customers

YANG Senbin

(College of Automation, Guangdong University of Technology, Guangzhou 525000, China)

Abstract: The application of data mining technology in the catering industry has a certain social value. Based on the prediction of the customer flow in the catering industry, restaurants can prepare meals for customers accordingly, which improves not only the customer experience of dining but also the quality of service as well as more efficient operation of the catering industry. By studying linear regression algorithm and random forest algorithm theory, this paper puts forward the idea of combining linear regression algorithm with random forest algorithm, and applies it to the prediction of the numbers of customers return visits in restaurants, and proves the rationality and feasibility of the idea through experiments. Compared with linear regression algorithm with random forest algorithm, the accuracy of the combined algorithm is increased by about 3.004% and 2.022% respectively. Better prediction results have been obtained than most of the previous research, which provides a new way of thinking for the application of data mining technology in the catering industry.

Keywords: data mining; linear regression; random forests; algorithm combination; catering industry

1 引言(Introduction)

数据挖掘运用相关的算法从大量的、不完全的、有噪声的、模糊的、随机的数据中提取正确的、有用的、未知的、综合的,以及用户感兴趣的知识,建立模型,用于决策支持的模型,提供预测性决策支持的方法、工具和过程^[1]。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中的信息的过程。随着互联网时代的到来和数据大爆发,数据挖掘技术普遍而且迫切地应用于各个领域,比如金融、电信、保险、医疗、餐饮等行业。它通过运用整理、分析、总结、推理等方法对大量地数据进行处理,从而对实际问题进行指导和分析,得出相关地预测结果,从而做出更加有利的决策^[2]。

本文主要针对餐饮行业对餐厅客流量进行预测。餐饮是我们日常生活的重要一部分,同时也伴随着巨大的行为数据产生。如果能更好地利用这些数据,运用数据挖掘技术对数据进行分析、预测,让餐饮行业能更合理地做出方案和决策,这有利于顾客用餐体验,提高餐饮质量,同时也可以让餐饮行业经营更好,商家获得更大的利润。笔者查阅相关资料得知,数据挖掘在餐饮领域运用的普遍性还有待提高,数据挖掘运用在餐饮行业的相关技术还有待提高。如以往有不少人研究的线性回归模型或随机森林模型运用在餐饮行业仍然存在准确性不够的不足。本文针对这一不足提出了一个新的解决方法,整合两个模型的优点,避开它们的缺点,融合

成一个新的模型,提高数据挖掘技术在餐饮行业应用的水平。

本文通过介绍线性回归和随机森林两种模型,运用数据挖掘工具对餐饮行业大量数据进行分析预测,对比线性回归、随机森林和两个模型融合后的新模型之间的预测结果,通过对比结果数据,从而得出合理的论断,融合后的模型在餐饮行业客流量预测中的效果更好,准确性更高。

2 算法介绍(Algorithm introduction)

2.1 线性回归模型简介^[3,4]

对于 $y_i, x_{i1}, x_{i2}, \dots, x_{ip}, y_1, y_2, \dots$, 其中 n 组结果值 $y_1, y_2, \dots, y_n$ 为取值为0或1的随机变量, $x_1, x_2, \dots, x_p$ 满足与 y_i 的关系为:

$$E(y_i) = \pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \quad (1)$$

式(1)中, $f(x)$ 是单调递增函数, 其值域为 $[0, 1]$, y_i 是均值为 $E(y_i) = \pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})$, 分布在0—1, 概率函数为:

$$P(y_i = 1) = \pi_i, P(y_i = 0) = 1 - \pi_i \quad (2)$$

y_i 的概率函数也可用式(3)表示:

$$P(y_i) = \pi_i y_i (1 - \pi_i) (1 - y_i) \quad y_i = 0, 1; i = 1, 2, \dots, n \quad (3)$$

所以, $y_1, y_2, \dots, y_n$ 的似然函数取对数后结果为:

$$\begin{aligned} \ln L &= \sum_{i=1}^n y_i \ln \pi_i + (1 + y_i) \ln(1 - \pi_i) \\ &= \sum_{i=1}^n [y_i \ln \frac{\pi_i}{(1 + \pi_i)} + \ln(1 - \pi_i)] \end{aligned} \quad (4)$$

然后, 将式(5)

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})} \quad (5)$$

代入(4)中, 得到线性回归:

$$\begin{aligned} \ln L &= \sum_{i=1}^n [y_i (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \\ &\quad - \ln[1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]] \end{aligned} \quad (6)$$

选取 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 的估值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$, 使式(6)达到最大值。

2.2 随机森林算法简介

随机森林(random forest)是一种基于分类树(classification tree)的算法(Breiman, 2001)。这个算法需要模拟和迭代, 被归类为机器学习中的一种方法。经典的机器学习模型是神经网络(Hopfield, 1982), 有半个多世纪的历史了。神经网络预测精确, 但是计算量很大。20世纪80年代Breiman等人(1984)发明了分类和回归树(Classification And Regression Tree, 简称CART)的算法, 通过反复二分数据进行分类或回归, 计算量大大降低^[5]。

RF是由一系列树型分类器 $\{h(x, \theta)\}_k$, 其中 $k=1, \dots, k$

合成的分类器, 其中 θ_k 是独立同分布随机向量, 且每棵树对输入向量 x 所属的最受欢迎类投一票^[6]。RF生成步骤如图1所示: (1)从总训练样本集 D 中用Bootstrap采样选取 k 个子训练样本集 $D_1, D_2, \dots, D_k$, 并预建 k 棵分类树; (2)在分类树的每个节点上随机地从 n 个指标中选取 m 个, 选取最优分割指标进行分割; (3)重复步骤(2)遍历预建的 k 棵分类树; (4)由 k 棵分类树形成随机森林。

Bootstrap随机抽样得到输入训练集和节点随机选取指标进行分割, 使得RF对噪声有很好的容忍性, 且降低了分类树之间的相关性。单棵树不剪枝任意生长的特点可获得低偏差分类树, 且能够保证对新测试数据分类的正确率。

RF的生成和单棵风险分类树如图1和图2所示^[7]。

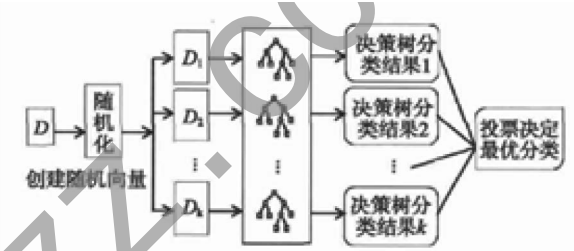


图1 RF生成步骤

Fig.1 RF generation step

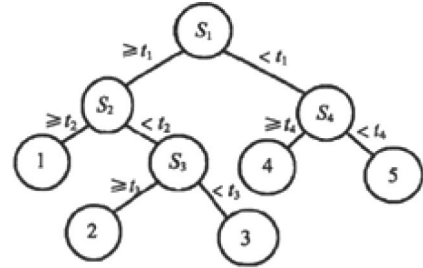


图2 单棵风险分类树

Fig.2 Single risk classification tree

2.3 线性回归和随机森林算法融合原理

本文使用的模型融合是加权平均法。首先, 根据两个模型的预测效果和模型得分, 对其求权重; 然后对两个模型预测值求加权平均值为最终的预测结果。

$$y_i = y_{i_LR} * a + y_{i_RFR} * b$$

其中, y_i 表示最终的预测结果, y_{i_LR} 表示线性回归模型的预测结果, y_{i_RFR} 表示线性回归模型的预测结果, a 表示求得的线性回归模型的权重, b 表示求得的随机森林模型的权重。

2.4 模型评价标准

对于客流量的预测, 我们更加关注的是所预测的客流量与实际客流量之间的误差^[8], 因此, 本文采用均方根误差RMLSE作为评价模型效果优良的标准。

其中， $RMLSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (observed_t - predicted_t)^2}$ ^[9]

$observed_t$ 表示真实访客数量， $predicted_t$ 表示预测的访客数量。

当RMLSE越小时，表明误差越小，模型效果越好。

3 模型数据分析(Data analysis of model)

3.1 数据预处理与特征工程

餐饮店客流量预测的数据来源与kaggle数据竞赛平台Recruit Restaurantor Visitor Forecasting赛题中的数据。本文经过数据清洗和预处理^[10]后提取出对本次实验有用的328298条记，每条记录12个字段，见表1。

表1 预处理后的属性记录

Tab.1 Property record after preprocessing

air_store_id,is_true,visit_date,visitors,day_of_week,is_holiday,prev_day_is_holiday,next_day_is_holiday,air_genre_name,air_area_name,latitude,longitude
预处理后的数据并不能满足实验的要求，仍然有大量的潜在特征价值未被挖掘，此时需要进行相应的特征工程，本次实验对访客数量做log处理、时序问题趋势稳定处理、关于顾客就餐是否为周末等时间问题处理等特征工程，充分挖掘了数据中隐藏的大量价值。特征处理后每条记录62个字段，如表2。

Tab.2 Characteristics after the feature engineering

air_store_id,is_true,visit_date,day_of_week,is_holiday,prev_day_is_holiday,next_day_is_holiday,air_genre_name,air_area_name,latitude,longitude,is_outliters,visitors_log,is_weekend,day_of_month,Day_of_week,vc_fewm_mean,vc_fewm_std,vc_log_fewm_mean,vc_log_fewm_std,vc_day_of_week_fewm_mean,vc_day_of_week_fewm_fewm_std,vc_log_day_of_week_fewm_mean,vc_log_day_of_week_fewm_fewm_std,vc_isholiday_fewm_mean,.....

表2 特征工程处理后的属性

3.2 模型构建

在对数据进行预处理和特征工程之后，就可以输入数据进行模型训练，其中除表2中的visitors_log属性外为输入数据集的X，visitors_log属性为标签输入。模型输入数据如表3所示。

表3 模型的数据输入

Tab.3 Data input of the model

air_store_id	is_true	visit_date	day_of_week	
air_00a91d42b08b08d9	False	2016-07-01	Friday	
air_00a91d42b08b08d9	False	2016-07-02	Saturday	
air_00a91d42b08b08d9	False	2016-07-03	Sunday	
air_00a91d42b08b08d9	False	2016-07-04	Monday	
air_00a91d42b08b08d9	False	2016-07-05	Tuesday	
.....				

表4为模型输出的示例，以Logistic回归作为预测模型。其中，Id表示店铺的地址和该店铺对应的日期，visitors表示该店铺在对应日期访客量的预测人数。

表4 模型的数据输出

Tab.4 Data output of the model

Id	visitors
air_00a91d42b08b08d9_2017-04-23	2.706023
air_00a91d42b08b08d9_2017-04-24	10.433438
air_00a91d42b08b08d9_2017-04-25	14.114471
air_00a91d42b08b08d9_2017-04-26	18.580581
air_00a91d42b08b08d9_2017-04-27	14.807312
.....	

3.3 实验结果与分析

3.3.1 各模型预测值与真实值对比

为了便于观察实验现象和分析实验结果，本文在数据规模一致且合理的情况下分别对线性回归模型、随机森林模型和两个模型融合后预测的访客量数与真实访客量数进行对比，并作图分析。由图3—图5分析可得，线性回归模型、随机森林模型和融合后的模型在一定程度上三个模型的预测值和真实值是吻合的。分析可得，三个模型适用于本文实验，模型应用合理。

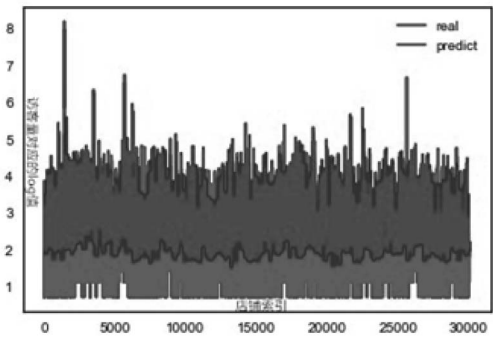


图3 线性回归模型预测值和真实值对比图

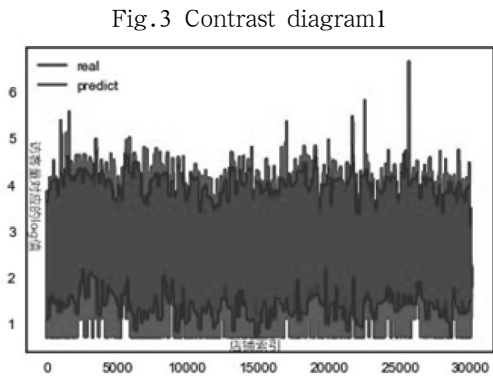


图4 随机森林模型预测值和真实值对比图

Fig.4 Contrast diagram2

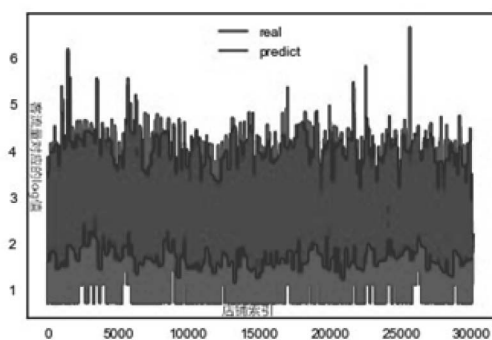


图5 融合模型预测值和真实值对比图

Fig.5 Contrast diagram3

3.3.2 不同数据集大小的情况下各模型误差对比

分别取全部数据的20%、40%、60%和100%，分别训练以上三个模型，得出三个不同模型的错误率，并作图对比，如图6所示。

其中，图6图例LR、RFR和Fusion model分别表示线性回归模型、随机森林模型和融合后的模型的错误率。

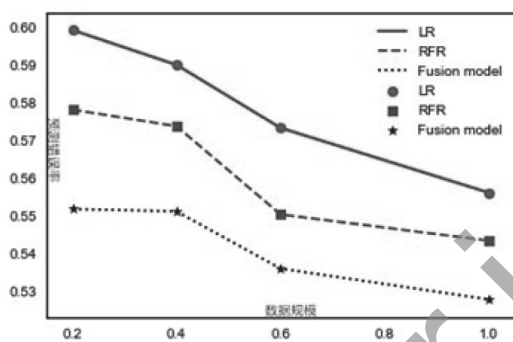


图6 预测误差率对比图

Fig.6 Prediction error rate contrast diagram

3.3.3 实验结果分析

由图3—图5分析可得，线性回归模型、随机森林模型和融合后的模型在一定程度上三个模型的预测值和真实值是吻合的。分析可得，三个模型适用于本文实验，模型应用合理。分析图6，对于某一模型，随着数据量不断增加，模型预测预测访客的错误率不断降低，模型效果不断提高，并且当数据量大到一定范围时，模型预测访客数量的错误率下降趋缓；对比三个模型可得，在相同的数据规模下，融合后的模型预测访客数量的错误率明显低于线性回归模型和随机森林模型。

4 结论(Conclusion)

本文介绍了线性回归和随机森林算法，以及两个算法

融合思路及原理，把数据挖掘技术运用于餐饮行业，并通过实验证明两个算法融合这一思路在餐饮行业预测餐厅某一段时间段访客数量的应用是可行合理的，具有较大的社会价值，让餐饮行业能更合理地做出方案和决策，这有利于顾客用餐体验，提高餐饮质量，同时也可以让餐饮行业经营更好，商家获得更大的利润。本文线性回归和随机森林算法融合思想弥补了之前单模型研究的不足，降低餐厅访客量预测的错误率，模型效果更优，使模型在餐饮行业使用范围更广，更容易泛化和推广；在一定范围内，随着训练数据规模增加，模型的效果不断优化。

参考文献(References)

- [1] 张晴,高广银,贾波.数据挖掘技术在超市营销系统中的应用[J].软件工程,2016,19(5):35-38.
- [2] 张慧,徐勇.数据挖掘中SVM模型与贝叶斯模型的分析——基于电信客户的流失分析[J].平顶山学报,2016,31(2):68-73.
- [3] 冷建飞,高旭,朱嘉平.多元线性回归统计预测模型的应用[J].统计与决策,2016,16(7):82-85.
- [4] 王朋,吕寒,李若诗.基于logistic回归的《中国好声音》受众行为分析[J].广州大学学报,2013,12(11):54-58.
- [5] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(4):1190-1197.
- [6] Jonathan D W,Jennifer L I,Suzana J C.Coastal flooding by tropical cyclones and sea-level rise[J].Nature,2013,504(8):44-52.
- [7] 赖成光,陈晓宏,赵仕威,等.基于随机森林的洪灾风险评价模型及其应用[J].水利学报,2015,46(1):58-66.
- [8] 刘伟,徐鹏涛.O2O电商平台在线点评有用性影响因素的识别研究——以餐饮行业O2O模式为例[J].中国管理科学,2016,24(5):168-176.
- [9] Dong L J,Li X B,Peng K.Prediction of rockburst classification using Random Forest[J].Transactions of Non-ferrous Metals Society of China,2013,23(2):472-477.
- [10] 李强,赵晨杰,罗先录.基于大数据应用技术的学情分析系统架构分析与设计[J].软件工程,2018,21(5):34-37.

作者简介:

杨森彬(1993—),男,硕士生.研究领域:数据挖掘,机器学习,文本挖掘.