Statistics and Application 统计学与应用, 2022, 11(2), 402-409
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112043

Hans汉斯

基于XGBoost算法的混合模型在零售业数据中的应用

王稼春，彭姜深，朱　莉

宁波工程学院，浙江　宁波

收稿日期：2022年3月24日；录用日期：2022年4月13日；发布日期：2022年4月24日

摘　要

本文使用德国零售数据，经过观察数据的特征，展开数据预处理，建立特征工程；通过将聚类分析与XGBoost模型相结合，建立混合预测模型对德国零售业数据进行建模预测。该模型首先对特征集降维之后选择最优的聚类数，然后对聚类分析后不同的类别分别进行XGBoost模型训练，最后将通过加权求和得到预测结果。研究结果表明，相比较于其他模型，混合模型提升了预测精度和泛化能力。

关键词

XGBoost，聚类分析，混合模型，预测

Application of Hybrid Model Based on XGBoost Algorithm in Retail Data

Jiachun Wang, Jiangshen Peng, Li Zhu

Ningbo University of Technology, Ningbo Zhejiang

Received: Mar. 24th, 2022; accepted: Apr. 13th, 2022; published: Apr. 24th, 2022

Abstract

The paper uses German retail sales data, after observing the characteristics of the data, carries out data preprocessing, and establishes feature engineering; by combining cluster analysis with XGBoost model, a hybrid prediction model is established to model and forecast German retail sales data. Firstly, the model selects the optimal cluster number after dimensionality reduction of feature set, then performs XGBoost model training for different categories after cluster analysis, and finally

文章引用: 王稼春, 彭姜深, 朱莉. 基于 XGBoost 算法的混合模型在零售业数据中的应用[J]. 统计学与应用, 2022, 11(2): 402-409. DOI: 10.12677/sa.2022.112043

obtains prediction results through weighted summation. The studies show that the hybrid model improves the prediction accuracy and generalization ability compared with other models.

Keywords

XGBoost, Cluster Analysis, Hybrid Model, Forecast

1. 引言

在改革开放的背景下，我国市场经济正在飞速发展，零售业的竞争也愈加激烈。各企业若想在众多企业中脱颖而出，需要做出相对准确的预测。好的预测不仅可以为企业解决存货堆积的麻烦，还可以为企业提供以后的需求走向，这样以便于进行物流准备以及人员排期等一些前期准备工作[1]。以及为后期的生产和采购计划的制定提供可靠依据。

市场各销售行业销量的预测有着重要的研究价值和意义，近年来国内外学者对这些问题进行了深入的研究，目前常用的方法有时间序列分析[2] [3] [4]、支持向量机[5] [6]、神经网络等方法[7] [8]，以及将不同方法结合建立混合模型[9]。但是不同模型针对不同的数据的特点各有优缺点，因此如何找到最适合零售行业销售数据的预测模型才是最令我们关心的问题。

欧洲的零售行业相对于我国的零售行业较为发达，其中最为发达的国家是德国。本文针对德国一大型超市的销售额数据，提出将聚类分析与 XGBoost 模型相结合的混合预测模型，通过与多元线性回归模型、随机森林回归模型、XGBoost 回归模型对比，进一步得到对销售额预测最精准，泛化能力最好的模型，并可将其推广于中国零售业销售额数据的预测中，为零售企业提供辅助性的决策。

2. XGBoost 算法

XGBoost (eXtreme Gradieni Boosting)，是改进的梯度提升学习算法，是 Boosting 中的一种方法。XGBoost 通过对损失函数进行二阶 Taylor 展开，在损失函数里增加正则项，用于控制模型的复杂度。GBDT (梯度决策树)算法是 Boosting 方法中的重要组成部分，其算法步骤为：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N),x_i\in\chi\in R^n,y_i\in\gamma\in R\}$；损失函数为：$L(y_i,f_{t-1}(x))$。

1) 初始化弱学习器：

$$f_0(x)=\arg\min\sum_{i=1}^{m}L(y_i,c) \tag{1}$$

2) 对迭代轮数 $t=1,2,3,\cdots,T$ 有：

a) 对样本 $i=1,2,\cdots,m$ 计算负梯度：

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{2}$$

b) 利用 $(x_i,r_{ti})(i=1,2,3,\cdots,m)$，可以拟合一棵 CART 回归树，从而得到第 t 棵回归树，其对应的叶子节点区域为 R_{t_j}，$j=1,2,3,\cdots,J$，其中 J 为回归树 t 的叶子节点的个数。

c) 对上述叶子区域，计算最佳拟合值：

$$C_{tj} = \arg\min \sum_{x_i \in R_{ij}} L\left(y_i, f_{t-1}\left(x_i\right)+c\right) \tag{3}$$

这样就得到了本轮的最优决策树拟合函数 $h_t\left(x\right)=\sum_{j=i}^{J} C_{tj} I\left(x \in R_{tj}\right)$。

d) 更新强学习器并得到其表达式：

$$f_t\left(x\right)= f_{t-1}\left(x\right)+\sum_{j=1}^{J} C_{tj} I\left(x \in R_{tj}\right) \tag{4}$$

XGBoost 算法是在 GBDT 的基础上对算法进行以下优化。

在 XGBoost 模型进行对应 t 次迭代之后，得到：

$$\widehat{y}_i^{(t)} = \widehat{y}_i^{(t-1)} + f_t\left(x_i\right) \tag{5}$$

此时，XGBoost 的目标函数为：

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t\left(x_i\right)\right)+\Omega\left(f_t\right) \tag{6}$$

其中 $\Omega\left(f\right)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2$。将目标函数在 $\widehat{y}_i^{(t-1)}$ 处进行二阶泰勒展开，

$$L^{(t)} \cong \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y}_i^{(t-1)} + g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right)\right]+\Omega\left(f_t\right) \tag{7}$$

其中

$$g_i = \partial_{\widehat{y}_i^{(t-1)}} l\left(y_i, \widehat{y}_i^{(t-1)}\right), h_i = \partial^2_{\widehat{y}_i^{(t-1)}} l\left(y_i, \widehat{y}_i^{(t-1)}\right) \tag{8}$$

去掉公式中的常数项后，可以得到：

$$\tilde{L}^{(t)} \cong \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]+\Omega\left(f_t\right) \tag{9}$$

加入正则项：

$$\tilde{L}^{(t)} \cong \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{10}$$

进一步可以转化为：

$$\tilde{L}^{(t)} = \sum_{j=1}^{T}\left[g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{11}$$

解出使目标函数最小的 ω，从而可以得到目标函数的最优解。XGBoost 算法学习出的模型更加简单，可以防止过拟合。

3. 数据分析

在对数据进行建模分析之前，都要对原始数据进行预处理，这一步是整体工作流程当中非常重要的一个环节。模型的预测是否准确很大程度上取决于训练预测模型时输入数据的质量，因此需要对原始数据进行探索性数据分析。对那些有价值的数据搭建特征工程，将数据转换为训练模型能够接受的数据形式，进一步地去训练预测模型。

3.1. 数据描述

本文所使用的数据来自 Kaggle 网站关于德国 Rossmann 零售超市销售额数据。数据包含 1115 家 Rossmann 连锁商店在 2013 年 1 月 1 日到 2015 年 7 月 31 日的销售数据。影响销售额的因素有很多，如商店是否促销，是否国家法定节假日，是否学校放假，最近竞争商店的距离，最近竞争商店开业的月份和年份，是否连续促销等因素。下面对数据集进行特征描述，见表 1 和表 2。

Table 1. Characteristics of training set and test set
表 1. 训练集、测试集特征

特征名称	含义	值
Store	商店号	数值，范围：1 到 1115
Day of Week	星期几	数值，范围：1 到 7
Date	时间	YY/MM/DD
Sales	销售额	数值
Customers	顾客数	数值
Open	是否开店	0：关店，1：开店
Promo	当天是否有促销	0：无促销，1：促销
State Holiday	假日	0：非假日，a：公共假日，b：复活节，c：圣诞节
School Holiday	学校假日	0：非假日，1：假日

Table 2. Data characteristics of store information
表 2. 商店信息数据特征

特征名称	含义	值
Store	商店号	数值分，范围：1 到 1115
Store Type	商店类型	分类：a、b、c、d 四类
Assortment	商店类别	分类：a 基础类、b 补充类、c 扩展类
Competition Distance	最近的竞争对手的距离	数值，单位：米
Competition Open Since [Month/Year]	最近竞争商店开业的月份和年份	1~12 月
Promo2	连续促销	0：否，1：连续促销
Promo Since [Year/Week]	开始连续促销的年份和日历周	2013、2014、2015 日历周
Promo Interval	连续促销重启月份	“Jan., Apr., Jul., Oct.”, “Feb., May., Aug., Nov.”, “Mar., Jun., Sept., Dec.”

3.2. 特征工程

原始数据的所有属性都用于统计建模是不切实际的，必须根据模型要求和属性特征值进行特征工程构建。

具体的实现过程为：将 Date 日期单独分开，转换为 Year、Month、Day、Week of Year；去掉没有价值的数据：去掉商店关闭时的数据和商店营业时销售额是 0 的数据；因为 Sales 的数额都太大，在数据拟合的过程中会造成很多不良的影响，所以对 Sales 进行取对数处理，即新增一个 Saleslog 特征；Store 商

店信息数据集里面的 Competition Distance 竞争者距离的空值用中位数填充；Store 商店信息数据集里的空值用 0 来填充；合并商店信息数据集和训练集；State Holiday 特征的取值有 0、a、b、c，因为 a、b、c 在建立模型的时候不可以识别，所以将 a、b、c 做独热编码，同时 Store Type 特征的取值 a、b、c、d 和 Assortment 特征的取值 a、b、c 做一样的处理；Promo Interval 特征的取值有三种，定义 promo (函数)，判断促销是否再进行。如果月份的英文简写在促销月份之内，意味着促销正在进行。

4. 模型建立

模型的建立包含以下基本流程，见图 1：

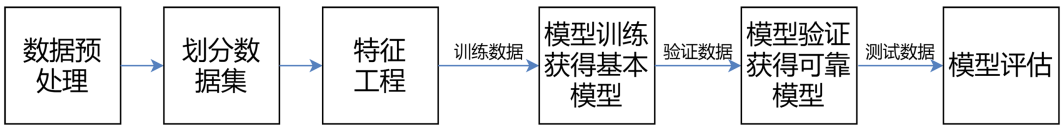

Figure 1. Basic process of model building
图 1. 模型构建的基本流程

1) 数据预处理：对原始数据进行一系列探索性分析，用探索性分析得出的结果对数据集进行预处理，其中包括异常值处理，空值填充，给特征工程提供数据。

2) 划分数据集：完成预处理后，把数据划分成训练集、测试集和验证集。

3) 特征工程：对训练集、测试集、验证集分别进行特征构建，生成相应的特征集，为接下来的构建模型做好准备。

4) 构建基本模型：分别选择线性回归模型、随机森林模型和 XGBoost 构建基本的销售额预测模型，同时对参数进行调优，通过调节一些关键参数的大小观察预测精度评价指标的变化，得到各模型的最优参数，建立基本的模型。

5) 模型优化及选择：对 XGBoost 进行改进及优化，对比各模型的预测结果，选出最佳预测效果的模型。

4.1. XGBoost 模型

本文利用 1115 家 Rossmann 连锁商店在 2013 年 1 月 1 日到 2015 年 7 月 31 日的销售数据进行建模预测。其中将 1,017,209 个数据作为训练集，50 个数据作为测试集。在实验中经过多次调试与测试，在权衡计算量与模型的综合得分后将 XGBoost 模型参数树的深度 max_depth 设置为 5，树的棵树 n_estimators 设置为 20，其余参数都设置为默认参数。探索 XGBoost 模型对销售额的预测性能，其实验结果如图 2 所示。

XGBoost 模型在预测中的 RMSPE (均方根百分比误差)为 0.0152，MAE (平均绝对误差)为 0.1042。该模型的预测效果较好。图 2 中可以清楚地看到预测效果虽然在趋势上有所接近实际销售额趋势，但是总体上销售额的预测值和真实值还存在较大偏差，因此该模型还需要改进。

对每个特征的重要程度评价打分，如图 3 所示，从打分情况可以看出“Customers”和“Competition Distance”的特征得分比较高，排名靠前，即这两个特征对模型非常重要。

4.2. 建立混合模型

从图 3 特征重要度得分图中得出“Assortment”和“Store Type”这两特征的重要程度得分都很低，对预测销售额的贡献不大，在影响销售额的众多特征中不重要。而根据探索性数据分析可以知道，“Store Type”与“Customers”、“Sales”和人均消费数都有较直接的关联，这就说明单一的 XGBoost 回归模型在构建模型的时候会损失“Assortment”和“Store Type”这两个维度的信息，从而导致预测模型没有达到最优。

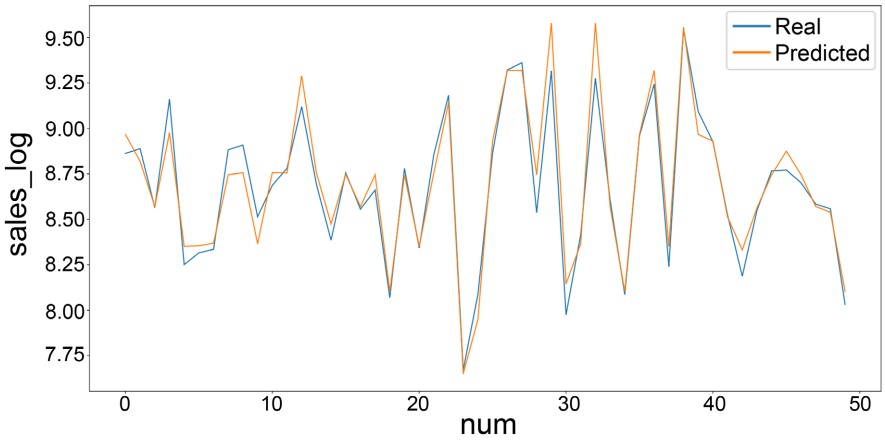

Figure 2. Sales forecast results of XGBoost model
图 2. XGBoost 模型销售额预测结果

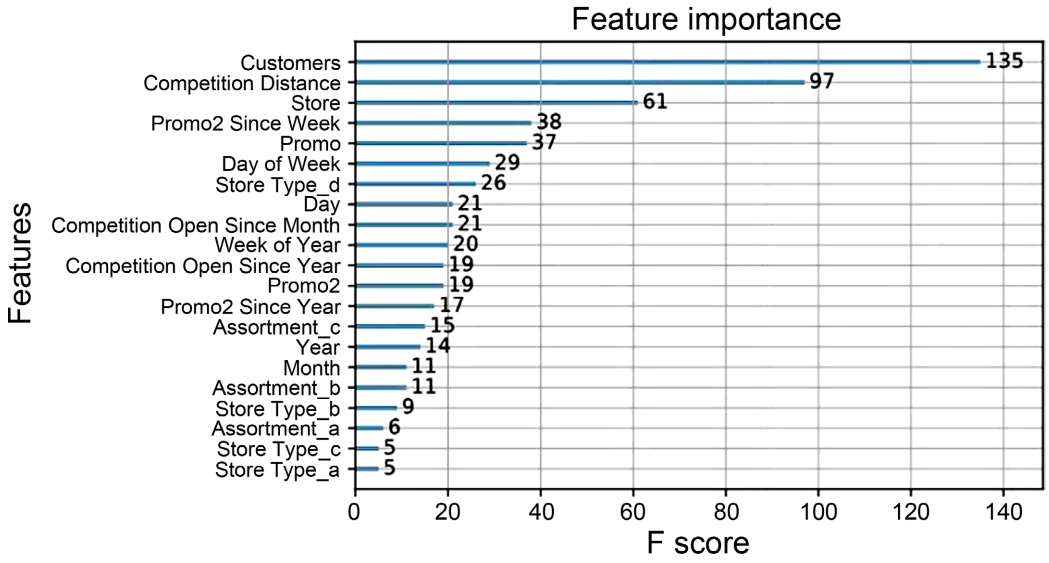

Figure 3. Significance score of features
图 3. 特征重要度得分

大型的商店具有地理位置各异、商店规模各异、消费群体各异等特点，若构建单一的模型不可以满足所有商店的销售额预测需求，因此本节尝试构建一个基于聚类方法的混合回归模型，通过这个来解决单一模型造成信息损失的问题。

混合模型利用主成分分析(PCA)的方法进行降维，按照贡献率来选取比较合适的降维维数。使用 K 均值聚类方法对数据进行聚类处理。对每一类数据分别建立 XGBoost 回归模型，最后将通过加权求和得到预测结果。

首先利用主成分分析对数据集的特征进行降维，如表 3 所示。可以发现当主成分个数是 3 的时候，累计贡献率达到了 0.99968912，因此提取 3 维特征矩阵来代替原来的特征矩阵。接下来进分别选择 K = 2, 3, 4, 5, 6, 7 的情况进行聚类，我们发现当 K = 5 时最合适。我们对聚类之后各类别的数据分别建立 XGBoost 模型，最终的预测结果见图 4。从图中可以看到，大部分真实值和预测值之间的差距变小了，基于 XGBoost 算法的混合模型在预测中的 RMSPE = 0.0682，MAE = 0.009，说明混合模型的拟合程度很高，模型的预测准确度也高，基本达到了对销售额预测的要求。

Table 3. Principal component contribution rate and cumulative contribution rate of all principal components
表 3. 主成分贡献率及所有主成分累计贡献率

主成分	贡献率	累计贡献率
1	0.46443623	0.46443623
2	0.36215851	0.82659474
3	0.17309438	0.99968912
4	0.00031007	0.99999919

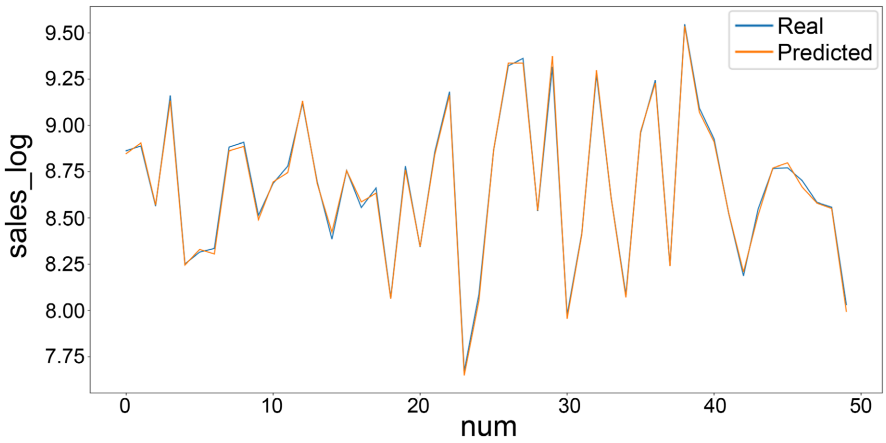

Figure 4. Prediction effect of mixed model
图 4. 混合模型预测效果

表 4 分别给出三种基本模型和混合模型在预测精度上的比较。在三种单一的模型中，XGBoost 模型在预测精度指标 MAE、RMSPE 中都是最小的，这说明 XGBoost 的拟合效果是最好的，预测是最精准的，因此选择 XGBoost 模型进一步建立混合模型。综合四种模型预测结果的对比，混合模型在两个精度评价指标上都是最小，表明混合模型的性能最好，精确度最高。

Table 4. Evaluation indexes of accuracy of different models
表 4. 不同模型精度的评价指标

模型	RMSPE	MAE
线性回归模型	0.1596	0.0247
随机森林回归模型	0.1431	0.0208
XGBoost 回归模型	0.1042	0.0152
混合回归模型	0.0682	0.0090

5. 结论

本文通过对德国 Rossmann 连锁商店的零售数据进行探索性分析，搭建特征工程，建立了 XGBoost 模型。为了优化模型，进一步建立了基于聚类分析和 XGBoost 算法的混合模型。通过将混合模型与三种单一模型在数据预测中的结果进行比较，我们发现混合模型进一步提高了预测的精度。本文所提出的模型不仅能应用在德国的零售企业中，也对我国的实体零售业以及线上电商平台的商品定价、运营方式都具有重要意义。

基金项目

国家级大学生创新创业训练计划项目(202111058036)，宁波市自然科学基金(2021J144)。

参考文献

[1] Chang, P.-C., Wang, Y.-W. and Liu, C.-H. (2007) The Development of a Weighted Evolving Fuzzy Neural Network for PCB Sales Forecasting. *Expert Systems with Applications*, **32**, 86-96. https://doi.org/10.1016/j.eswa.2005.11.021

[2] 陈宇科. 商品销量的趋势分析及预测[J]. 渝西学院学报(自然科学版), 2003(2): 59-61.

[3] Wu, L., Yan, J.Y., Fan, Y.J. (2012) Data Mining Algorithms and Statistical Analysis for Sales Data Forecast. 2012 *Fifth International Joint Conference on Computational Sciences and Optimization*, Harbin, 23-26 June 2012, 577-581. https://doi.org/10.1109/CSO.2012.132

[4] 牟书成. 面向零售业时间序列预测与分析的算法研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019

[5] 杜小芳, 张金隆. 农产品销量预测的支持向量机方法[J]. 中国管理科学, 2005(4): 129-134.

[6] 武牧, 等. 一种基于支持向量机的卷烟销量预测方法[J]. 烟草科技, 2016, 49(2): 87-91.

[7] Qin, Y.Q. and Li, H.M. (2011) Sales Forecast Based on BP Neural Network. 2011 *IEEE* 3*rd International Conference on Communication Software and Networks*, Xi'an, 27-29 May 2011, 186-189. https://doi.org/10.1109/ICCSN.2011.6014419

[8] 马超群, 王晓峰. 基于 LSTM 网络模型的菜品销量预测[J]. 现代计算机(专业版), 2018(23): 26-30.

[9] 张凌波, 刘海. 基于 IF0A-SVR 的断路器销量预测[J]. 控制与决策, 2019, 34(12): 2667-2672.