Hans Journal of Data Mining 数据挖掘, 2018, 8(2), 74-78
Published Online April 2018 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2018.82008

Hans汉斯

Supermarket Commodity Sales Forecast Based on Data Mining

Yanmei Jiang, Qingkai Bu

Qingdao University School of Electronic Information, Qingdao Shandong
Email: jymdoit@126.com

Received: Apr. 9th, 2018; accepted: Apr. 20th, 2018; published: Apr. 27th, 2018

Abstract

Based on the comparison of several basic models, a prediction model based on LightGBM and support vector regression model is proposed in this paper. This model not only extracts the features of the user's behavior data and the features of commodity attributes, but also combined with the advantages of time sliding window in feature processing, extracts dynamic features by using the sale data of the commodity and correlation data, and then uses the fusion of multiple models to predict the commodity data. The experimental results show that after the feature extraction of sliding window method, by comparing support vector regression model and LightGBM prediction model, it is found that the effect of LightGBM prediction model is slightly better than the support vector regression model. By combining the support vector regression model and the LightGBM model, the root-mean-square error of the supermarket sales forecast model is 1.23209, which is significantly higher than the single model prediction results. Therefore, this model is an effective method to predict the sales volume of short-term supermarket.

Keywords

LightGBM, Dynamic Feature Extraction, SVR, Model Combination

基于数据挖掘的超市商品销量预测

姜艳梅，卜庆凯

青岛大学电子信息学院，山东 青岛
Email: jymdoit@126.com

收稿日期：2018年4月9日；录用日期：2018年4月20日；发布日期：2018年4月27日

摘 要

针对超市商品短时间内销量预测问题，本文通过对比几种基本模型，提出了一种基于LightGBM和支持向

文章引用: 姜艳梅, 卜庆凯. 基于数据挖掘的超市商品销量预测[J]. 数据挖掘, 2018, 8(2): 74-78.
DOI: 10.12677/hjdm.2018.82008

量回归模型相结合的预测模型。该模型不仅通过对用户的行为数据进行量化特征提取和商品属性的特征提取，同时结合了时间滑动窗口在特征处理上的优势，将商品的销售数据作为前后关联数据进行动态特征提取，再通过多模型关系的融合，对商品数据进行预测。实验结果显示，经过滑窗法特征提取后，通过对比支持向量回归模型和LightGBM预测模型，发现LightGBM预测模型效果略优于支持向量回归模型，通过组合支持向量回归模型和LightGBM模型，发现超市销量预测模型的均方根误差仅为1.23209，明显高于单模型预测结果。因此，该模型是短期超市商品销量预测的一种有效方法。

关键词

LightGBM，动态特征提取，SVR，模型组合

1. 引言

现如今传统行业的竞争愈演愈烈，超市的形态也更加多样化，智能化，若要在大数据的营销环境中，争取一席之地，必须将数据化技术引入其中，建立与客户的密切联系，实现精准的销售营销策略[1]。目前，可用于购买行为预测的模型与方法很多，包括决策树方法，贝叶斯分类算法、支持向量机、神经网络方法以及时间序列预测方法等[2]。预测的方法和预测的数据不同，产生的效果也不同。预测方法各有其优点和缺点，但都是对数据从不同角度来进行的解读，因此各个模型的相互组合能够发挥不同方法的优势，对数据有全面综合的理解。目前很多企业已经将数据挖掘与传统的预测方法相结合，极大的改变了超市销售预测的局面[3]。许多大型零售商倾向于使用时间序列方法进行预测。这需要大量的销售数据作为数据支撑。对于数据量较少或者小型零售商，则倾向于用使用经验法定性分析[4]。

对于时间序列预测模型而言，本文所采用的数据为短期销售数据，时间跨度小，不能很好的体现销量的周期性变化。支持向量机回归(SVR)是根据结构风险最小化原则提出的[5]，具有很好的泛化能力。因此针对短期数据的特点，本文通过滑动时间窗口增加时间序列特征，并采用支持向量回归模型进行动态特征提取，通过组合模型，最终建立合理有效的预测模型，从而进行准确的销量预测。该模型较单一的模型具有更好的稳定性和更高的预测准确率。

2. 基本模型概述

支持向量回归模型介绍

支持向量回归模型其核心思想是通过引入非线性映射 $\varphi(x)$，将原始的低维特征空间映射到高维的特征空间，在高维特征空间中构造最优分类超平面[5]。假设 (x_i, y_i)， $x \in R^d$， $y_i \in R$， $i = 1, \cdots, n$，目标是求解下列回归函数

$$f(x) = (w \cdot \Phi(x)) + b$$

其中，w 是权值向量，x 表示模型的输入，b 是误差值，而 $\Phi(x)$ 则表示核函数。对于样本 (x, y)，传统的回归模型通常是基于模型输出 $f(x)$ 与 y 的值的差值来计算损失支持向量回归[6]的一般形式可表示为：

$$f(x, w) = f(x, \alpha, \hat{\alpha}) = \sum_{i=1}^{n} (\alpha_i - \hat{\alpha}_i) K(x, x_i) + b$$

$K\left(x_i, x_j\right)$被称为核函数。只要符合 Mercer 条件的函数均可用作核函数。尽管可供选择的核函数很多，其中最广泛使用的核函数是径向基函数(RBF)。径向基函数可由下述方程得到：

$$K\left(x_i, x_j\right) = \exp\left(-\left\|x_i - x_j\right\|^2\right) \Big/ 2\sigma^2$$

其中σ表示 RBF 的宽度。本文选择高斯径向基函数作为核函数[7]，让 C 和ε在一定的范围内取值，将取定的 C 和ε不断用于训练集，通过网格搜索法最终使得训练集拟合程度最好的一组参数即作为最优参数。

LightGBM 模型介绍

LightGBM 是一个梯度提升框架，使用基于学习算法的决策树[8]。该算法是对随机森林的进一步改进，在模型的树模型中包含了分类树和回归树[9]。决策树常用来处理分类问题，在商品销量预测中可以对商品的离散性特征进行有效的处理和预测，回归树常用来处理预测问题，对商品的时间等连续性特性更加敏感。LightGBM 采用梯度提升的方式[10]，将分类树和回归树进行有效的叠加，该算法在对商品销量预测中，可以有效的将商品的基本属性，如类别，周期性指数等离散体征与按时间滑动窗口获取的连续销量的连续特征有效的结合，使得商品销量预测的多方面特征有一个更加综合的使用。

3. SVR 与 LightGBM 相结合的购买行为预测模型

本次设计所采用的数据是取自于 2015 年某城市的超市日常交易数据，该超市在一个地市级的小连锁超市，数据为 1~7 月份完整的交易数据。本文取 1~7 月份的 79,116 条交易记录作为训练数据。被购买过的商品类别有 849 种，中类类别有 189 种。

本文模型构建主要分为以下阶段，见图 1。

3.1. 预处理

预处理阶段主要完成两个工作，一是清理数据中的冗余数据和脏数据；二是提取商品某一段时间的销售行为。由于超市数据的特殊性，逻辑数据不存在问题，但易出现人员操作导致数据错误问题。根据现有资料及数据自身特点，用相关数据进行填写[11]。对于商品某段时间的销售行为，选取 7 天，30 天分别为一个时间窗口，根据用户编号和商品编号对窗口期的数据进行提取，形成以时间为顺序的用户对特定商品的购买行为数据[12]。

3.2. 训练集和验证集

由于数据量的限制，在做数据处理事，利用滑窗法增加数据量，构造 3 个数据集进行验证，分别如下表 1 所示。

3.3. 特征工程构建

征体系由消费者特征组，商品特征组，其他特征组组成。特征提取阶段是对预处理之后的购买行为数据和购买行为对应的商品进行。该过程分为如下 2 个阶段：

1) 静态特征提取：对商品的特征提取是根据购买行为特征数据中商品编号[12]，从已有的商品信息中提取该商品的类别信息、功能分类信息和商品的属性信息。商品的特征数据与购买行为特征数据组合，形成用户购买行为的静态特征。在商品销售特征行为中，分别对预测期前 7 天，30 天的商品的销量，回购量，售卖人数等求最大值、最小值、和值、均值等的统计值[13]，来作为用户近一段时间的商品行为特征。

2) 动态特征提取：根据对商品销量趋势数据[14]的分析，发现 95%以上的商品在时间窗口内的销量

大于 10。通过将商品的基本特征和销售特征作为训练数据，利用时间滑动窗口，用 SVR 模型进行处理，提取商品销售的商品销量周期性指数，用来作为商品销量的动态特征。

4. 实验结果与评价

4.1. 评价标准

预测回归类预测模型精度评价的方法[15]常用的有平均绝对误差(mean absolute error, MAE)，均方根误差(root mean squared error, RMSE)，平均百分比误差(mean percentage error, MPE)和平均绝对百分比误差(mean absolute percentage error, MAPE)。本文选择 RMSE 来作为判断标准，RMSE 的大小表示预测值与真实值之间的差异 ，RMSE 值越小，模型最后预测结果的精度就越高。预测模型精度等级分类见表 2 所示。

4.2. 实验结果

实验中分别采用随机森林、支持向量回归模型，梯度提升回归模型上文中的数据进行预测。实验显示，对于支持向量回归模型，惩罚系数 C = 0.1，损失函数中的 ε 参数 = 10，rmse = 2.50119595807；对于 LightGBM，feature_fraction' = 0.8，'bagging_fraction': = 0.8，leaves = 24，learning_rate = 0.3，rmse = 2.236。对比发现，单一模型的情况下，支持向量回归模型的模型效果，次于 LightGBM。说明 LightGBM 在处理连续性特征和离散性特征的效果更优于 SVR 模型。通过对 SVR 和随机森林的模型融合，得到的 rmse = 1.23209。说明较单一模型而言，组合模型能更好的发挥各模型的优势，较单一模型有更好的提升。

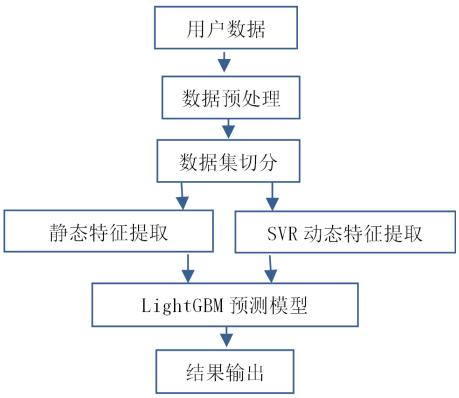

Figure 1. Flowchart
图 1. 流程图

Table 1. Segmentation data set by sliding window method
表 1. 滑窗法切分数据集

数据表示	特征集	预测集
Submit1	4,5,6,7 月份数据	8 月份
Submit2	3,4,5,6 月份数据	7 月份
Submit3	1,3,4,5 月份数据	6 月份

Table 2. Evaluation level
表 2. 评价等级

等级	好	较好	合格	不合格
均方根误差	<1.0	2.0~1.0	3.0~2.0	>3.0

5. 结束语

在对超市商品销量预测的过程中，先前发表的论文大多是基于长时间的销量预测，采用时间序列进行分析。本文针对短时域的销量预测，创新的采用动态特征提取的方式，通过模型组合的方式，证明以动态特征为基础的组合模型的预测效果，明显高于各单一模型的预测。但预测结果仍受社会因素，促销活动，天气状况等非线性变量因素的影响。接下来的实验，通过利用聚类分析，挖掘商品销售的潜在分类，研究其对销量的影响。

参考文献

[1] 朱明. 数据挖掘[M]. 北京: 中国科学技术大学出版社, 2002.

[2] 冯萍, 宣慧玉. 数据挖掘技术及其在营销中的应用[J]. 食品科学技术学报, 2001, 19(1): 52-58.

[3] Jiawei Han, Micheline Kambe. 数据挖掘: 概念与技术[M]. 机械工业出版社, 2012, 21(3): 105-106.

[4] 赵改平, 刘丽兰, 程功勋, 树志松. 销售预测分析系统的研究与应用[J]. 现代制造工程, 2011(3): 28-31.

[5] 李斌, 鄱涛, 史明华, 等. 基于支持向量机的交通流组合预测模型[J]. 天津工业大学学报, 2008, 27(2): 73-76.

[6] 杨金芳, 翟永杰, 王东风, 徐大平. 基于支持向量回归的时间序列预测[J]. 中国电机工程学报, 2005, 25(17): 110-114.

[7] 侯振雨, 蔡文生, 邵学广, 等. 主成分分析-支持向量回归建模方法及应用研究[J]. 分析化学, 2006, 34(5): 617-620.

[8] Wang, D., Zhang, Y. and Zhao, Y. (2017) LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients. *Proceedings of the* 2017 *International Conference on Computational Biology and Bioinformatics*, Newark, NJ, 18-20 October 2017, 7-11. https://doi.org/10.1145/3155077.3155079

[9] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1): 137-141.

[10] 王焱. 基于随机梯度提升决策树的行人检测算法设计与实现[D]: [硕士学位论文]. 杭州: 浙江大学, 2017.

[11] 王丰效, 郭天印. 季节性商品销售量预测模型[J]. 陕西工学院学报, 2003, 19(2): 84-87.

[12] Cappellari, L. and Jenkins, S.P. (2003) Multivariate Probit Regression Using Simulated Maximum Likelihood. *The Stata Journal*, **3**, 278-294.

[13] 戴亮, 孟晶. 零售业销售预测方法[J]. 商场现代化, 2011(1): 83-85.

[14] 叶倩怡. 基于 Xgboost 方法的实体零售业销售额预测研究[D]: [硕士学位论文]. 南昌: 南昌大学, 2016.

[15] 张婧. 基于数据挖掘的零售业商品销售预测研究[D]: [硕士学位论文]. 成都: 四川师范大学, 2008.