计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.008919]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多模式特征聚合的未来商业预测①

崔铭浩, 张仁博, 郭恩铭

(东北大学 计算机科学与工程学院, 沈阳 110169)
通信作者: 崔铭浩, E-mail: cmhstuedu@163.com

摘　要: 准确预测商业销售量未来趋势对于企业开发经营、政府宏观调控等至关重要. 传统的数据预测方法计算时间开销大, 具有主观性, 而现有基于数据驱动的未来商业预测方法没有考虑到数据集中的特征多样. 商业销售量数据是一个时序数据, 时序数据中包含了丰富的时间窗特征、滞后历史特征和价格变化趋势特征等众多特征, 先前的研究往往只注重于其中的某些特征, 对于特征的融合和增强探究偏少, 现有的未来商业预测方法的预测精度仍然有待提高. 为此, 本文提出了一种基于多模式特征聚合的未来商业预测方法, 该方法首先将商业销售量数据进行预处理; 然后基于特征工程提取数据集的 5 组不同的时间窗特征和其他特征; 在机器学习上对于 5 组时间窗特征采用硬投票机制选择合适的模型训练, 同时也采用神经网络的优化模型提取时序特征和预测结果, 然后分析销售量数据集和某些特征之间的依赖关系; 最后基于软投票模型完整地模型融合实现了商业销售量的高精度预测. 一系列实验结果表明, 本文提出的方法具有较高预测精度和效率, 明显优于现有预测方法.

关键词: 未来商业预测; 多模式; 特征融合; 投票机制; 机器学习; 深度学习

引用格式: 崔铭浩,张仁博,郭恩铭.基于多模式特征聚合的未来商业预测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/8919.html

Future Business Forecasting Based on Multi-mode Feature Aggregation

CUI Ming-Hao, ZHANG Ren-Bo, GUO En-Ming

(School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

Abstract: Accurate prediction of the future trend of the commercial sales volume is of great importance to the development and operation of enterprises and the macro-control by the government. Traditional data prediction methods are time-consuming and subjective, while the existing data-driven future business prediction methods do not take into account the diversity of features in the data sets. The data of the commercial sales volume is time-series, which contains a wealth of time window features, lagging historical features, and price change trend features. Previous studies tend to focus only on some of these features, and the integration and enhancement of these features are seldom explored. The prediction accuracy of the existing future business prediction methods still needs to be improved. Therefore, this study proposes a future business forecast method based on multimodal feature aggregation, which firstly preprocesses the commercial sales volume data and then extracts five different groups of time window features and other features of the data set on the basis of feature engineering. In machine learning, the hard voting mechanism is used to select the appropriate model for the training of the five groups of time window features. At the same time, the neural network optimization model is applied to extract the time-series features and forecast results, and then, the dependency relationships between the data set of the sales volume and some features are analyzed. Finally, with the soft voting model, a high-precision forecast of the commercial sales volume is achieved by complete model integration. The experimental results reveal that the proposed method has high prediction accuracy and efficiency, which is greatly better than the existing prediction methods.

Key words: future business forecasting; multi-mode; feature fusion; voting mechanism; machine learning; deep learning

① 收稿时间: 2022-06-01; 修改时间: 2022-07-01; 采用时间: 2022-07-24; csa 在线出版时间: 2022-10-28

准确预测商业销售量未来趋势对于企业开发经营、政府宏观调控等至关重要. 传统的数据预测方法计算时间开销大, 具有主观性, 而现有基于数据驱动的未来商业预测方法没有考虑到数据集中的特征多样. 商业销售量数据是一个时序数据, 时序数据中包含了丰富的时间窗特征、滞后历史特征和价格变化趋势特征等众多特征, 先前的研究往往只注重于其中的某些特征, 对于特征的融合和增强探究偏少, 现有的未来商业预测方法的预测精度仍然有待提高. 为此, 本文提出了一种基于多模式特征聚合的未来商业预测方法, 该方法首先将商业销售量数据进行预处理; 然后基于特征工程提取数据集的 5 组不同的时间窗特征和其他特征; 在机器学习上对于 5 组时间窗特征采用硬投票机制选择合适的模型训练, 同时也采用神经网络的优化模型提取时序特征和预测结果, 然后分析销售量数据集和某些特征之间的依赖关系; 最后基于软投票模型完整地模型融合实现了商业销售量的高精度预测. 一系列实验结果表明, 本文提出的方法具有较高预测精度和效率, 明显优于现有预测方法.

1 引言

当今世界的经济迅速发展, 也带来了竞争国际化[1]. 在商业预测中, 销售量和价格是众多企业关注的, 但近两年的新冠疫情的变化, 国内国外的电子商务市场在其运行过程中却一直处在不断的波动中[2], 众多市场的大起大落, 使商业现状变化更加急峻. 一个企业一旦发生了经营内外环境的变化, 那么如何做出一个很好的决策和计划主要取决于科学的预测, 同时科学的预测依赖于企业对于经济和商业市场变化规律的认识, 也依赖于企业对于预测技术的了解和熟练程度, 然后让企业根据科学的预测结果做出相关的干预措施, 比如确定经营目标, 制订销售决策和生产计划, 这有助于企业未来的发展. 因此, 对商业其中的一些标准衡量预测, 避免非正常波动的出现, 既可为企业开发经营、投资决策提供依据, 又可为政府宏观调控、制定政策提供参考[3,4].

商业预测方法可大致分为两大类, 一类是定性分析方法, 另一类是定量分析方法[5]. 定性分析方法中有经验判断预测方法和特尔菲法, 前者通过个人和集体基于经验进行判断, 容易受到各种心理因素的影响, 后者是通过匿名方式结合专家的意见对发展做出量的推断, 但是缺乏数据支持, 仅代表大体方向. 定量分析方法中有基于时间序列预测方法的模型[6], 也有基于回归分析预测法的模型[7], 在商业趋势稳定发展且价格没有大幅度浮动的情况下, 前者可以较好地预测未来趋势, 后者是广为使用的预测模型, 但是采取的特征过多或者过少的时候会出现拟合异常现象, 导致预测效果不是很佳.

因此本文采用了每一个模型的多个特征进行组合起来, 形成大量的相关特征组, 提出一种基于特征融合的未来商业预测方法. 该方法首先将商业销售量数据进行预处理, 在预处理过程中提出了一个滑动窗口特征, 依次划分 5 组不同大小和步长的窗口, 然后利用机器学习中的多个模型进行训练, 同时采用神经网络的优化模型收集时间特征融合到销售量的特征组中; 然后对于每一个模型基于投票机制将从 5 组窗口选择出较优的模型; 最后基于模型融合利用这些较优的模型得到的预测进行时间序列模型的预测, 从而预测出来商业未来趋势. 通过实验验证本文相较于传统方法中单一模型预测有较好的预测效果, 首先是在窗口组大小的设置, 对于每一个模型来说都选择了较优的窗口大小, 其次基于软投票机制完成了对模型的预测结果的融合, 在通过机器学习与深度学习中的单个模型预测和投票模型融合预测的对比之下在准确率评价标准下数值增加了 2.31%, 误差值降低了 1.46%.

2 基于特征融合的商业未来预测

针对未来商品销售量的预测, 本文提出了一种多模式特征聚合的商业预测方法, 主要由 3 部分构成, 分别为数据预处理、特征选择和处理、模型融合预测, 方法总体框架如图 1 所示, 同时关于在特征选择和处理的环节上, 对于投票模块的框架如图 2 所示.

2.1 数据搜索

数据探索在特征工程是非常重要的一个环节, 面对数据探索, 首先需要对数据集进行一个宏观分析, 包括数据的缺失和重复、异常值检测以及一些数据的清洗工作[8], 然后本文基于数据的特征要进行对其相互关系的分析, 包括计算相关性, 变量可视化等, 在进行数据探索中, 本文可以进行一系列的特征选择, 通过预处理可以清洗构造出新的数据集, 为特征工程的第一步做出很重要的贡献.

在日常生活中, 本文经常遇见很多相关的数据类型, 比如数值、分类变量、文本和图像的数据, 图像数

据比其他的数据更复杂, 需要进行关键点的定位和方向确定[9], 在其他的数据的处理中可以采用归一化、标准化以及离散化等相关的算法[10] 进行筛选数据, 把有效数据进行特征选择, 才能有利于模型的选择和优化.

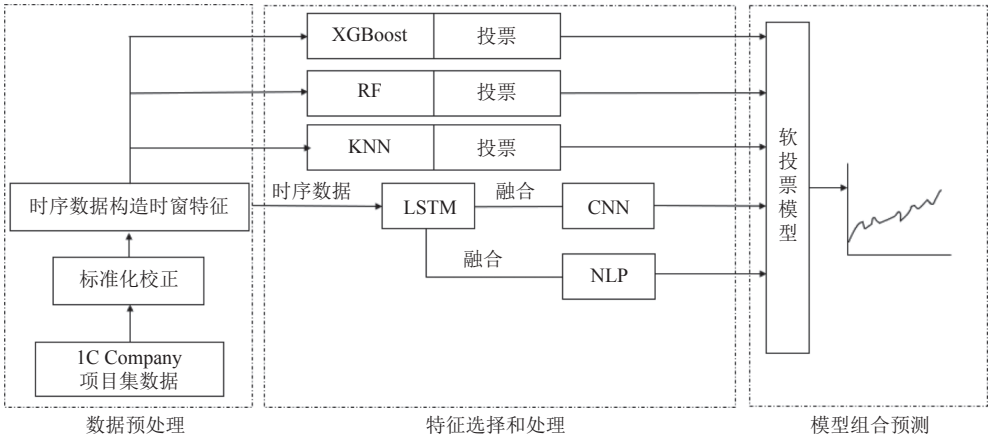

图 1 未来商业预测方法总体框架

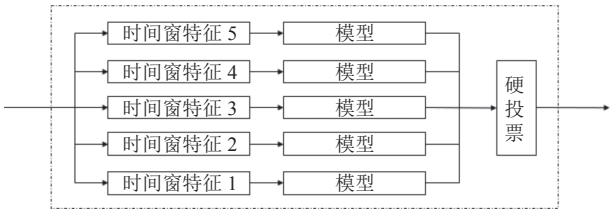

图 2 投票模块框架

2.2 特征构造

特征工程方法[11] 为电子商务产品销售量的某些特征的选择提供了一些参考依据, 在特征构造的时候, 除了最常见的统计量特征之外本文还需要做出其他相关的特征如时间特征.

假设某电子商务中商店种类有m个, 商品种类有n个, 本文可以构造一个二维向量X, 对于商品每一个种类来说, 本文可以提供k个特征进行选择, 而模型训练的选择将由这些特征所影响, 本文将对每一个特征进行模型训练, 这样本文对特征的选择面会扩大, 这样在使用模型预测的时候, 才具有更多的预测可能性和可信性.

在未来销售量预测中, 本文以商店信息和商品信息形成有序对, 首先按照月份上计算出销量和价格的总值特征item_cnt、transactions和均值特征mean_item_cnt, 按照月份、商品种类和商店来分组构造与销量相关的关系曲线. 其次本文要构造label特征, 在数据搜索上已经处理数据集相关年月信息, 同时为了满足项目数据集的预测要求在时间上后移, 即一月份的label特征信息是二月份的销量. 然后本文构造了单位价格特征unit和产品的价格波动特征increase、decrease等相关的特征进行对单位价格的变化的趋势进行描述.

为了让模型训练的效果更具有可选择性, 本文的关键点在于构造时间窗特征[12], 其窗口window和步数step本文将划分成 5 组, 每一组对应不同的窗口大小和步数大小, 构造 min、max、mean、std 特征, 因为时间窗可以对数据起到平滑的效果, 同时也包含了一些相关的历史信息, 这样方便本文可以去构造新的滞后历史特征trend. 本文使用滞后平移操作和时间窗的计算划分验证集和数据集, 因为所给数据没有对应的验证集, 因此本文没有用以往的交叉验证[13] 去划分处理, 采用了特征组合[14] 的方法.

本文根据时间窗特征中的窗口大小依次划分训练集和验证集, 为商店、商品、年和月构造销量的均值特征shop_mean、item_mean、shop_item_mean、year_mean和month_mean, 与此同时为了对测试集缺失特征的填充, 因此构造了一个记录特征lastest_records, 该特征表示如果每验证集的最后一个月出现了某种商店信息和商品信息的组合, 则该组合记录一定是在最后一个月, 如果没有则寻找最近的一个月的特征记录并填充.

2.3 模型训练

本文在进行模型训练之前, 首先是特征的选择. 由于对于每一个模型都有自己的特征重要性排序, 因此这极大方便本文为每一个模型选择合适的特征, 并进行了模型训练. 在选择模型的时候, 本文从特征构造的角度出发, 采用了机器学习和深度学习的多种模型进

行训练. 在机器学习上本文采取了 XGBoost 模型、随机森林模型和 KNN 模型进行训练, 并且采用了投票机制对较好特征的选择. 在深度学习上本文采取了 LSTM 模型去训练, 并在此 LSTM 的训练基础上分别进行优化成编码解码结构[15] 和卷积层结构[16], 均取得不错的训练效果.

XGBoost[17] 是 Boosting 算法的一种实现. XGBoost 是一种加法模型, 它包含K个基学习器, 每个基学习器会对第i个输入样本x_i进行预测, 然后把每个预测结果$f_k(x_i)$相加, 作为最终的输出$\hat{y}_i$.

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

XGBoost 的目标函数由损失函数l和正则项Ω组成 (其中$l(y_i,\hat{y}_i)$是损失函数, $\Omega(f_k)$为正则项):

$$L(\phi) = \sum_i l(y_i,\hat{y}_i) + \sum_k \Omega(f_k) \tag{2}$$

对于树节点分裂来说, 精确算法是遍历所有特征的所有可能分割点, 如图 3 所示去寻找让目标函数最小的分割点, 近似算法则是对于每一个特征, 只考察分位点, 减少计算负责度, 但是对于 XGBoost 来说是以二阶导数值作为权重.

对于 XGBoost 模型来说, 本文重点利用 5 组不同窗口大小和不同步长大小的特征依次进行训练, 每一次训练都会形成一组预测结果和实验评估指标, 本文在实验评价环节上分析该指标.

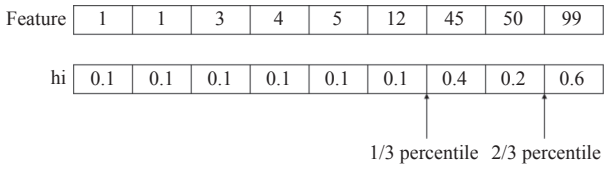

图 3 分割条件

很多算法其实覆盖了集成学习的思想, 随机森林是其中一个算法, 通过多棵树的集成随机组合, 它的基本单元就是决策树. 每一个决策树都是一个分类起, 那么对于一个输入样本, N棵树会有N个分类结果[18]. 而随机森林集成了所有分类投票的结果, 将投票次数最多的类别指定为最终的输出, 同时在树的每一个节点处, 从特征维度中随机选择m个特征维度 (其中m<<特征维度), 使用这些被选择的特征维度的最佳特征去分割节点. 在森林生长期间, m值保持不变, 接着让每一棵树都尽最大可能去生长, 并且忽略掉剪枝的过程[19].

无监督聚类模型中最具标志性的是 K-means 算法[20], 其最终的效果就是将数据分成一个个的簇, 本文采用该算法进行对时间滑动窗口的五组特征进行分析选择出较好的时间窗特征.

首先随机生成几个点, 成为聚类中心, 中心的个数一般和数据的类的种类数相关. 接着使用迭代算法: 先进行簇分配, 接着是移动聚类中心. 簇分配就是遍历数据集里面的每一个数据, 然后根据每一个数据距离各个聚类中心的距离最小来进行分配, 直到聚类中心不再变化, 即每一个数据被分配到的聚类不变. 根据每一个聚类的均值, 本文都要将聚类中心移动到这个点处, 在这个部分中, 如果出现没有聚类中心的现象, 这个时候需要移除这个聚类中心所在的聚类或者重新选取初始的聚类中心. 在进行分类的时候, 要注意对目标的优化、初始化选取的数据点而防止局部最优[20] 等.

投票机制[21] 是一种遵循少数服从多数原则的集成学习模型, 通过多个模型的集成降低方差, 从而提高模型的鲁棒性和泛化能力, 对于硬投票机制来说, 每个算法的预测都被认为是选择具有最高票数的类的集合, 对于软投票机制来说对各类测结果的概率进行求和, 最终选取概率之和最大的类标签, 换句话说硬投票是结果标签的集成, 软投票是概率的集成. 软投票法考虑到了预测概率这一额外的信息, 因此可以得出比硬投票法更加准确的预测结果, 当投票集合中使用的模型能预测类别的概率时, 适合使用软投票, 软投票同样可以用于那些本身并不预测类成员概率的模型, 只要它们可以输出类似于概率的预测分数值. 本文对前 3 大模型对构造的时间窗特征组合的选择采取了硬投票机制, 划分预测结果范围, 对其概率进行求和, 最终选取了预测概率较高的时间窗特征, 然后对于每一个模型的预测结果采用软投票机制进行模型融合.

LSTM 的组成部分包括遗忘门、输入门或更新门, 输出门. 遗忘门是 LSTM 单元的关键组成部分, 可以控制信息的保留与遗忘, 并且以某种方式避免了梯度消失和梯度爆炸的问题; 输入门用于控制网络当前输入数据图片流入记忆单元的多少, 即有保留保存到图片中的一部分信息; 输出门控制记忆单元图片对当前输出图片的影响, 即记忆单元中的某一部分会从时间步t输出, 下面是关于 3 个门的计算公式:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4}$$

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{5}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \times \tanh(C_t) \tag{7}$$

同时本文进行模型进行优化采用了编码-解码结构 (LSTM-MLP), 编码器基本上由一系列 LSTM 或 GRU 单元组成, 接受输入序列并输出每个时间步长的隐藏状态, 图 4 所示中的h_1, h_2, h_3, h_4并将该信息封装为内部状态向量. 解码器如图 5 所示使用编码器的输出和内部状态即由一个时间步长偏移的整个目标序列, 并输出每个时间步长 (例如s_1, s_2, s_3) 的隐藏状态.

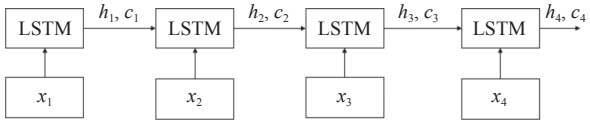

图 4 编码器结构

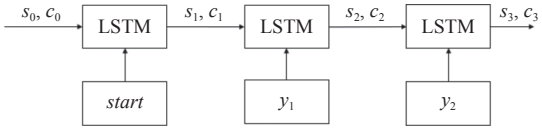

图 5 解码器结构

除此之外, 本文还采取了膨胀卷积结构 (LSTM-CNN). 该卷积结构在卷积时的输入存在间隔采样, 而采样率受图 6 的d控制. 最下面一层的表示输入时每个点都采样, 中间层表示输入时每 2 个点采样一个作为输入, 越高的层级使用的d越大. 所以, 膨胀卷积使得有效窗口的大小随着层数呈指数型增长. 更广泛的视野感受仅用层数较少的卷积网络就可以轻松获得.

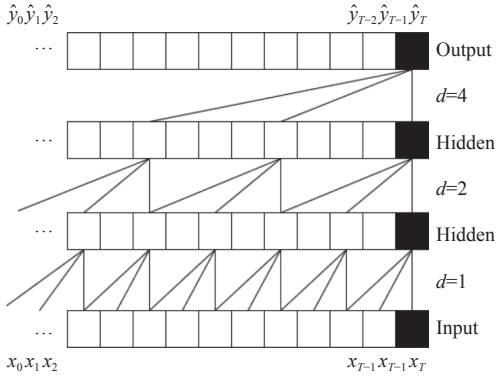

图 6 膨胀卷积结构

2.4 模型融合

本文在第 1 层建立许多模型, 但是现状多数关于时间序列模型的建立是针对于单个模型, 对于多个模型进行融合很少实现, 而且每一个模型的预测角度是不一样的, 因此本文基于投票机制将多个模型的预测结果融合.

本文记录众多模型预测出来的结果, 对于模型融合, 本文需要做的操作是比对各个模型的 *MSE* 以及其他相关性能分析, 然后采用软投票机制完成最后的模型融合最后一个环节.

软投票机制也称加权平均概率投票, 是使用输出类概率分类的投票法, 其通过权重得到整个融合模型预测结果的加权平均值, 而权重值是根据每一个模型的 *MSE* 来权衡计算过来的, 计算过程如式 (8) 所示:

$$level\_pred = \frac{\sum_i^n w_i \times machine_{pred_i}}{n} \tag{8}$$

3 实验过程

3.1 实验数据

本文所采用的实验数据来自于 Kaggle 项目中的未来商品销售量预测的数据, 训练集有 293 万行, 记录了每天每个店铺每种商品的单价和销量等 10 个相关特征, 测试集以商店和商品作为有序对, 共有 21.4 万行, 数据时间范围从 2013 年 1 月 1 日–2015 年 10 月 31 日, 相邻两次数据间隔一天.

3.2 实验环境

本文实验所采用的软硬件环境配置如表 1 所示.

3.3 实验分析

本文首先对数据进行预处理, 融合了 5 组窗口大小不同的时间窗特征, 将采取到的众多特征对其进行了相关性分析如图 7 所示. 通过相关性分析可知, 除商店商品数量、商品价格外, 相关程度较高的特征均集中在时间窗特征, 而且时间窗特征涉及到价格的变化趋势特征, 因此本文重点关注模型训练过程中对 5 组时间窗特征的选择.

表 1 实验环境配置

环境		配置
硬件环境	CPU	Intel(R) Xeon(R) CPU E5-2620 v4 2. 10 GHz
	GPU	NVIDI RTX 2080 TI
	内存	64 GB
软件环境	操作系统	Ubuntu 16.04.6
	开发语言	Python 3.7.6
	开发框架	PyTorch 1.4.0
	开发工具	PyCharm 2019.2.3 Professional

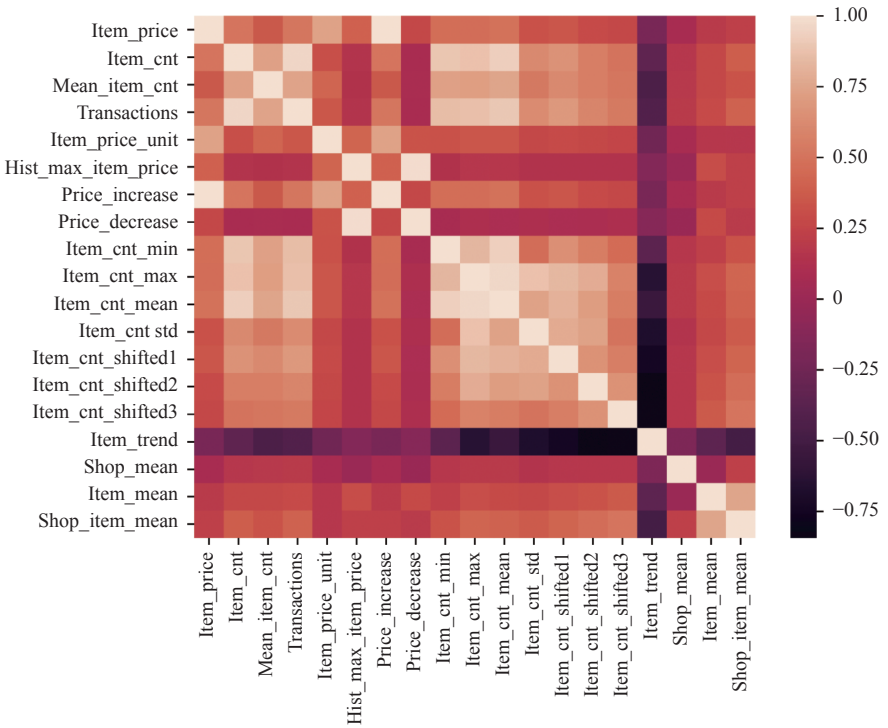

图 7　特征热图关联矩阵

与此同时在模型的特征重要性分析如图 8 所示, 本文发现到时间窗特征和滞后历史特征对于销售量的重要性程度比较高, 说明价格波动对销售量具有一定的影响, 同时商品的价格和数量的均值特征也排在前列, 因此本文可以将措施的重点安排在价格的宏观调度上.

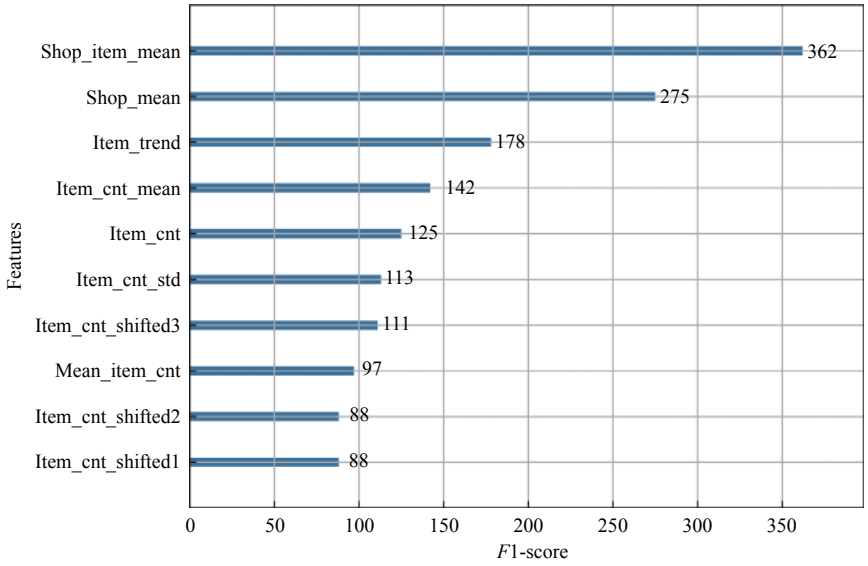

图 8　模型特征重要性分析

对于 5 组时间窗特征的选择, 本文在模型训练的时候利用硬投票机制选择合适的时间窗特征, 在投票过程中本文采用分数表 (表 2) 来比较每个算法的精度分数分布, 本文采用分层 10 倍交叉验证 3 次重复的分数来表示模型训练分数. 由模型分数可以知道对于 XGBoost 模型来说选择第 5 组时间窗特征进行训练比

较好, 对于 RF 模型来说选择第 4 组时间窗特征进行训练比较好, 对于 KNN 模型选择第 3 组时间窗特征比较好.

表 2 模型得分数

时间窗特征	XGBoost	RF	KNN
$shift1$	0.873	0.875	0.879
$shift2$	0.889	0.879	0.884
$shift3$	0.895	0.885	0.904
$shift4$	0.899	0.902	0.894
$shift5$	0.900	0.891	0.897

本文采用均方误差 (*MSE*) 评估指标来评估所提出模型的性能. 假设预测值$\hat{y}=\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$, 真实值$y=y_1,y_2,\cdots,y_n$, *MSE* 的计算公式见式 (9):

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \tag{9}$$

MSE 的范围为$[0,+\infty)$, *MSE* 越大, 表示预测值和真实值相距越远, 即模型的拟合效果越差; *MSE* 为 0, 则表示模型的预测结果和真实结果完全一致.

本文同时采用了 MAPE 和 R^2 来评估模型的性能, 如果 MAPE 越小, R^2 越接近于 1, 模型的性能越好.

在神经网络进行时间序列模型预测的时候, 模型的训练过程已经为本文构造相关时间特征, 然后本文采用了两种优化方式的模型进行预测, 都对原有模型进行 4 种预测步长下的对比, 如图 9、图 10 所示. 由图可以清楚知道在每一种预测步长下, 传统 LSTM 方法其 *MSE* 值高于两种优化模型, 这说明加入优化结构的 LSTM 模型具有很好的预测精度, 同时可以看出随着预测步长的增加, 优化前后的模型的预测误差也在增大, 这和理论预期相一致, 同时随着训练次数增加, 每一个模型的 *MSE* 都有所降低, 但是仍有优化模型的 *MSE* 低于传统模型的 *MSE*, 因此可以作为后期软投票的模型融合的一部分.

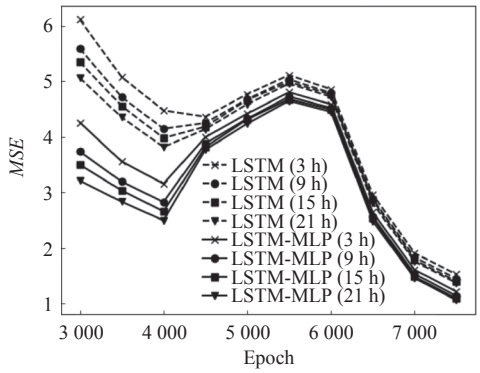

图 9 优化模型一 *MSE* 随训练次数的变化情况

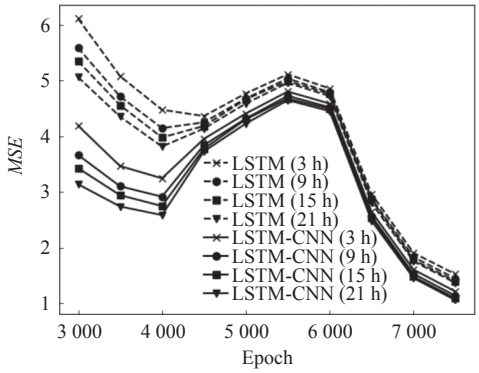

图 10 优化模型二 *MSE* 随训练次数的变化情况

本文将对 5 个模型进行比较相关的 *MSE*、MAPE 和 R^2 的值 (表 3), 发现 5 个模型的 *MSE*、MAPE 和 R^2 均相似, 说明每一个模型的预测结果也是相似同时较准确的, 首先不可否认每一个模型的性能都特别突出, 预测效果会比较优秀, 因此本文想通过一种融合的机制, 去验证融合后的模型的准确率是否有所提升, 从而进一步提高对于预测结果的准确率, 于是通过软投票机制将去融合其预测结果, 提出了一种投票融合的预测方案.

表 3 模型评估指标比较

模型名称	*MSE*	MAPE	R^2
XGBoost	0.8993	0.074	0.7152
RF	0.8985	0.067	0.7201
KNN	0.8846	0.072	0.7197
LSTM-CNN	0.9014	0.064	0.7048
LSTM-MLP	0.8945	0.068	0.7213

与此同时, 通过样品准确量和样品总量进行除法运算来表示模型的准确率. 我们发现通过将神经网络、XGBoost、RF 和 KNN 加入到投票组合中, 将对单个模型进行预测分析做对比 (表 4), 软投票准确率为 89.93%, 比最佳个体算法 XGBoost 模型的准确率 87.62% 提高了 2.31 %, 在模型的特异度和敏感度的比较上, 软投票模型是最高的, 说明该模型辨别的能力是高于其他模型的, 与此同时对于 *F*1 得分情况, 软投票模型最高, 同时通过得分的平衡权值, 可以看到除软投票模型之外的 LSTM-CNN 模型的得分是最高的为 84%, 与软投票模型相比, 误差值降低了 1.46%. 因此本文的软投票模型是可行的.

图 11 是商品未来销售量预测的结果, 在对单个模型进行预测的时候出现个别点突增的现象, 比如在某一个样品出现商品在当月销售量高达 14 个, 而在进行

投票融合之后的预测模型, 这中和了这些个别点突增的现象, 这样造成的波动就会很小, 因此本文的预测比较稳定, 不会出现很大的突变情况. 通过测试结果本文可以知道虽然这些销量的波动程度的变化趋势不是很突兀, 但是这些变化离不开政府企业的调控和预防[22–24].

表 4 预测结果评估指标

模型名称	Accuracy (%)	Sensitivity	Specificity	*F*1-score
XGBoost	87.62	0.8021	0.7941	0.7981
RF	86.53	0.7825	0.8021	0.7921
KNN	86.94	0.7912	0.7863	0.7887
LSTM-CNN	87.54	0.8495	0.8505	0.8397
LSTM-MLP	87.47	0.8120	0.8429	0.8271
投票组合	89.93	0.8461	0.8626	0.8543

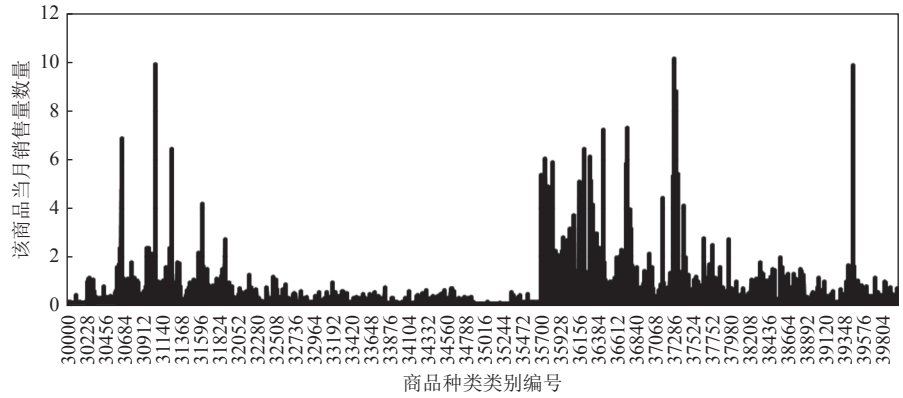

图 11 商品未来销售量预测结果

本文针对于其他文章的预测模型[25–29] 对本次实验进行综合分析进行对比, 如表 5 所示, 发现对于投票组合之后的模型融合预测方案较现有的文章的准确率还是偏高的.

表 5 现有文章其他算法对比

现有文献	模型内容	准确率 (%)	机器学习	神经网络	模型融合
[25]	KNN聚类和机器学习回归算法	84.63	√	—	—
[26]	基于优化的BP神经网络算法	85.97	—	√	—
[27]	Lasso回归算法和ARIMA模型	86.04	√	—	—
[28]	改进的人工神经网络算法	85.84	—	√	√
[29]	熵权法的Stacking算法集成且利用LightGBM模型进行融合	87.52	√	—	√
本文	采用XGBoost、RF、KNN和优化的神经网络进行投票组合	89.93	√	√	√

本文经过一系列的实验分析与比较, 最后得到了未来商品销售量趋势的预测, 同时也可以看出来本次模型的融合比较成功, 同时与其他单个算法比较, 本预测具有一定的准确度.

4 结论与展望

商品的相关特征对商业预测具有重要的意义, 对于企业发展、政府调控等具有重要影响, 因此准确预测商业未来销售量趋势对于相关企业和政府具有非常重要的意义. 本文结合先进的机器学习技术和神经网络技术, 提出了一种基于多模式特征聚合的商业未来销售量趋势的方法, 该方法首先将商业销售量的相关数据进行预处理, 然后利用特征工程多广度地提取多组时间窗特征, 并采用硬投票机制进行特征选择, 在此数据基础上采用两种优化的神经网络模型 LSTM-MLP 和 LSTM-CNN 时序数据预测结果, 实现了单个模型的预测结果, 基于模型结果的相似性, 本文采用软投票机制将众多模型进行融合, 从而预测出商业未来销售量趋势. 一系列真实数据上的实验结果表明, 本文提出的基于特征融合的未来商业预测方法, 明显优于现有的单个模型的预测结果, 同时对于时间窗特征的选择, 采用硬投票机制取得了较好的预测精度, 后续将扩大模型的应用范围, 进一步优化模型的结构, 进一步提升模型的预测精度和效率.

参考文献

1 李广春. 用新发展理念破解发展难题. 瞭望, 2021, (10): 1–2.

2 王可山, 郝裕, 秦如月. 农业高质量发展、交易制度变迁与网购农产品消费促进——兼论新冠肺炎疫情对生鲜电商发展的影响. 经济与管理研究, 2020, 41(4): 21–31.
3 徐德顺. 后疫情时代中国跨境电商发展的政策建议. 对外经贸实务, 2021, (7): 4–7. [doi: 10.3969/j.issn.1003-5559.2021.07.001]
4 陈一孚, 李贵武. “互联网+”驱动产业创新机制、商业模式与政策建议. 科学管理研究, 2021, 39(4): 87–91.
5 姜妍. 商业分析预测模式解析. 商业经济, 2020, (4): 118–119. [doi: 10.3969/j.issn.1009-6043.2020.04.047]
6 李文中, 万晨, 张治杰, 等. 一种基于自演化预训练的多变量时间序列预测方法和设备: 中国, 202010876972.2. 2020-11-17.
7 刘灿灿, 徐明瑜, 陈佳欣. 回归分析在并购重组评估实践中的应用——以价值比率的选取为例. 中国资产评估, 2022, (2): 4–11. [doi: 10.3969/j.issn.1007-0265.2022.02.001]
8 高菲, 宋韶旭, 王建民. 多区间速度约束下的时序数据清洗方法. 软件学报, 2021, 32(3): 689–711. [doi: 10.13328/j.cnki.jos.006176]
9 郭伟, 庞晨. 改进生成式对抗网络的图像数据集增强算法. 电讯技术, 2022, 62(3): 28–287. [doi: 10.3969/j.issn.1001-893x.2022.03.001]
10 张玉霖. 基于主成分分析的网络时延特征数据提取仿真. 计算机仿真, 2020, 37(3): 301–304. [doi: 10.3969/j.issn.1006-9348.2020.03.063]
11 谢斌, 林珊玲, 林志贤, 等. 基于强化学习的特征工程算法研究. 电子技术应用, 2021, 47(7): 29–32, 43.
12 田朝霞, 张俊, 陈旭, 等. 基于滑动时间窗的稠密子图发现算法研究. 计算机应用与软件, 2021, 38(7): 302–309. [doi: 10.3969/j.issn.1000-386x.2021.07.047]
13 张欣怡, 袁宏俊. 正则化和交叉验证在组合预测模型中的应用. 计算机系统应用, 2020, 29(4): 18–23. [doi: 10.15888/j.cnki.csa.007254]
14 李世奇, 赵铁军, 李晗静, 等. 基于特征组合的中文语义角色标注. 软件学报, 2011, 22(2): 222–232. [doi: 10.3724/SP.J.1001.2011.03844]
15 徐少峰, 潘文韬, 熊赟, 等. 基于结构感知双编码器的代码注释自动生成. 计算机工程, 2020, 46(2): 304–308, 314. [doi: 10.19678/j.issn.1000-3428.0053873]
16 张卫, 刘宇红, 张荣芬. 可实现时分复用的 CNN 卷积层和池化层 IP 核设计. 计算机工程与应用, 2020, 56(24): 66–71. [doi: 10.3778/j.issn.1002-8331.2003-0035]
17 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. 785–794.
18 吴地尧, 章新友, 张玉娇, 等. 分类算法在中药研究中的应用及其进展. 科学技术与工程, 2019, 19(35): 1–9. [doi: 10.3969/j.issn.1671-1815.2019.35.001]
19 Hutter F, Kotthoff L, Vanschoren J. Automated machine learning. Berlin: Saint Philip Street Press, 2020. 9–13.
20 章宦记. 改良的 K-means 与 K 近邻算法特性分析. 电子产品世界, 2016, 23(1): 79–80. [doi: 10.3969/j.issn.1005-5517.2016.01.025]
21 周翔, 翟俊海, 黄雅婕, 等. 基于随机森林和投票机制的大数据样例选择算法. 计算机应用, 2021, 41(1): 74–80.
22 张梦霞, 蒋国海. 政府短期消费刺激政策对经济复苏的作用机制研究——基于发达国家与发展中国家比较的多案例诠释. 财经问题研究, 2022, (2): 24–32. [doi: 10.19654/j.cnki.cjwtyj.2022.02.003]
23 刘敏. 政府治理能力与经济增长的门槛效应——以“一带一路”沿线国家为例. 经济问题探索, 2020, (1): 128–137.
24 陈大鹏, 吴舒钰, 李稻葵. 中国构建开放型经济的经验和对新发展阶段的启示——政府与市场经济学的视角. 国际经济评论, 2021, (6): 141–160.
25 周雨, 段永瑞. 基于聚类与机器学习的零售商品销量预测. 计算机系统应用, 2021, 30(11): 188–194. [doi: 10.15888/j.cnki.csa.008147]
26 孙艳文, 詹天明. 基于优化 BP 神经网络的销售预测算法研究. 计算机技术与发展, 2022, 32(1): 35–39. [doi: 10.3969/j.issn.1673-629X.2022.01.007]
27 房妮, 俱国鹏, 惠姣姣, 等. 基于 Lasso 回归和 ARIMA 模型的城市生活垃圾产生量预测——以宝鸡市为例. 河南科学, 2022, 40(1): 98–103. [doi: 10.3969/j.issn.1004-3918.2022.01.015]
28 姚兰, 戎荷婷, 褚超, 等. 基于人工神经网络的工业供应链销售预测方法. 计算机与数字工程, 2021, 49(10): 2057–2061, 2144. [doi: 10.3969/j.issn.1672-9722.2021.10.021]
29 张雷东, 王嵩, 李冬梅, 等. 多种算法融合的产品销售预测模型应用. 计算机系统应用, 2020, 29(9): 244–248. [doi: 10.15888/j.cnki.csa.007550]

(校对责编: 牛欣悦)