文章编号：2095-2163(2024)01-0175-05 中图分类号：TP391 文献标志码：A

基于改进注意力机制 Transformer 网络的快消品销量预测方法

王 阳，何利力，郑军红
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：销量预测能为企业生产计划、仓储运输提供决策支持，使企业能更好地适应市场需求。快消品销售量受众多因素的影响，具有季节性和周期性规律，传统的线性模型难以准确的预测，本文从长时序列预测的视角，运用深度学习理论，提出了一种基于订单时序和订单频率的改进自注意力机制模型（Sequence-Frequency Transformer，SFTransformer）。首先，基于快消品订单数据构建原始数据集，采用 time2vec 编码处理订单时序信息，并融合订单数据的时序和频率特征在基于时序的订单数据的不同订单频率分别对应不同的注意力头来关注订单数据的订单时序特征和频率特征；使用 Transformer 模型架构提取特征进行长时序列预测。在真实数据集上进行对比实验，SFTransformer 模型在均方误差（MSE）、平均绝对误差（MAE）、均方根误差（RMSE）3 项指标上均取得了最佳性能，验证了本文所提方法的有效性。

关键词：销量预测；长时序列预测；SFTransformer；改进自注意力机制

Cigarette sales prediction based on improved multi-head self-attention transformer

WANG Yang，HE Lili，ZHENG Junhong
（School of Computer Science and Technology，Zhejiang Sci-Tech University，Hangzhou 310018，China）

Abstract：Sales forecasting can provide decision support for enterprise production planning，warehousing，and transportation，enabling companies to better adapt to market demand. The sales volume of fast-moving consumer goods is influenced by numerous factors（content of factors），and exhibits seasonal and cyclical patterns. Due to the limitations of traditional linear models in accurately predicting sales，this paper proposes an improved self-attention mechanism model based on order sequence and order frequency from the perspective of long time series forecasting，using deep learning theory（Sequence-Frequency Transformer，SFTransformer）. Firstly，the original dataset is constructed based on fast-moving consumer goods order data，and time2vec encoding is used to process the order sequence information，integrating the sequence and frequency features of order data. Different order frequencies correspond to different attention heads to focus on the order sequence and order frequency features of the order data. Finally，the Transformer model architecture is used to extract features for long time series forecasting，and comparative experiments are conducted on real datasets. Experimental results demonstrate that the SFTransformer model achieves the best performance in terms of mean squared error（MSE），mean absolute error（MAE），and root mean square error（RMSE），validating the effectiveness of the proposed approach.

Key words：sales forecasting；long time-series prediction；SFTransformer；improved self-attention

0 引 言

随着大数据技术的迅速发展，生产企业的大量快消品的订单数据资源得以有效保留，这些数据资源蕴含生产企业的市场营销信息，大多数企业往往无法高效率低成本地运用销量数据来辅助商业决策。如何利用这些销量数据指导企业的生产决策仍是快消品生产企业面临的重要问题。通常来讲，快消品的销量与天气、交通运输量、股指期货等相似，都是一个时间序列的预测问题[1]。

时间序列的预测方法可大致分为传统统计方法、传统机器学习方法和深度学习方法。传统统计方法主要采用统计学知识对时间序列中蕴含的发展过程、方向和趋势进行建模并预测，常见的模型有自

基金项目：浙江省重点研发计划（2022C01238）。
作者简介：王 阳（1996-），男，硕士研究生，主要研究方向：数据智能；何利力（1966-），男，博士，教授、博士生导师，主要研究方向：数据分析、企业智能。
通讯作者：郑军红（1978-），男，博士，讲师，主要研究方向：商务智能、人工智能。Email：zdzhengjh@sohu.com
收稿日期：2023-01-28

回归(Auto Regressive, AR)模型、移动平均模型(Moving Average,MA)等。然而此类方法较低的表达能力不能处理复杂数据中的宏观趋势关系和非线性关系,预测准确率比较有限。传统机器学习方法包括支持向量机(Support Vector Machine,SVM)、贝叶斯网络等,时间序列预测效果良好。但由于快消品销售数据存在季节性、动态性、周期性及行业本身的特殊特征,数据序列往往存在很多干扰项,用传统机器学习方法较难进行精准的预测。

深度学习技术可以将有效特征从大量的原始数据中抽取出来,因此通过深度学习技术而建立的新模型实用性更强、准确度更高。循环神经网络(Recurrent Neural Network, RNN)由于具有记忆性、参数共享并且图灵完备(Turing completeness),在对序列的非线性特征进行学习时具有一定优势。循环神经网络在自然语言处理(Natural Language Processing, NLP),如语音识别、语言建模、机器翻译等领域有应用,也被用于时间序列预测领域。基于双阶段注意力机制的 RNN 模型(DA-RNN)是经典的时序预测模型[2]。DA-RNN 模型引入了 attention 机制,在编码的时候自适应选择相关程度高的序列信息进行编码,在解码阶段考虑编码阶段所有时间步骤的隐状态,而非传统方法的定长向量,从而解决长期依赖问题。长短期记忆(LSTM)网络是一类特殊的循环神经网络(RNN),能够学习长期依赖关系,LSTM 网络改善了 RNN 网络中存在的长期依赖问题,同时也更好地解决了 RNN 网络在训练过程中产生的梯度方面的问题。LSTM 可以在很长时间内保持状态中的时间信息,并广泛用于单变量和多变量领域的序列数据分析、预测和分类。门控循环单元(Gated Recurrent Unit,GRU)是 LSTM 的简化版本,将忘记门和输入门结合形成一个更新门,因此比 LSTM 具有更少的参数,但降低了复杂性,单元状态和隐藏状态也组合在一起,并使用重置门[3]。2017 年谷歌团队[4]提出的 Transformer 模型架构在长期预测问题中取得了不错的效果,Transformer 模型对于序列数据中的长期依赖关系显示出了强大的建模能力,非常适合于时间序列建模。为了解决时间序列建模中的特殊挑战,许多 Transformer 变体如 Informer、fastformer 等成功应用于各种时间序列任务中。

尽管关于长时序列预测已经有很多研究,但是针对快消品的长时段销售预测仍存在以下问题:

(1)在原始数据的处理中,未考虑快消品的高周转率和低保质期带来的高订单频次问题;

(2)传统模型更多考虑到订单数据的时序和销量对销售周期的影响,而忽略了订单频率中蕴含的细节信息。为解决以上问题,本文从长时序列预测的视角,运用深度学习理论,提出了一种基于订单时序和订单频率的改进注意力机制,并结合 Transformer 模型设计了时序-频率分解模型(sequence-frequency transformer, SFTransformer)。在模型中设计了时序-频率多头自注意力机制,基于时序的订单数据的不同订单频率分别对应不同的注意力头来关注订单数据的订单时序特征和频率特征。通过 time2vec 方法处理输入编码,最后使用 Transformer 模型架构提取特征进行长时序列预测。在真实数据集上进行对比的实验证明,本模型在实际的企业订单数据的预测上取得了较好效果,实验结果证明本文提出的 SFTransformer 模型在一定程度上提升了预测准确率,说明了模型的有效性。

1 模型构建

1.1 问题定义

本文提出的快消品订单销量预测问题定义:设滑动窗口定长为 N,预测步长为 k,模型输入第 i 个序列为离散时间序列 $T_i = \{x_{i+1}, x_{i+2}, \cdots, x_{i+N-1-k}, 0, \cdots, 0 \mid x_i \in R^{d_x}\}$,输出结果是预测的相应序列 $O_i = \{y_{i+1}, y_{i+2}, \cdots, y_{i+N-1} \mid y_i \in R^{d_y}\}$,其中 d_x 为输入的隐藏维度,d_y 为输出的隐藏维度。

1.2 输入编码

SFTransformer 模型的输入编码包括 3 部分:订单销量值编码 $data_{emb}$、订单频率编码 fre_{emb}、订单时间序列编码 seq_{emb}。为了使模型同时注意输入序列的订单时序信息和订单频率信息,除订单销量值外,还要将对应的订单时间序列和订单频率序列也融合到输入序列中,因此可选取时间序列的年、季、月、周等信息,分别对其进行编码并与订单销量值编码、订单频率编码进行融合。通过编码融合方法获取原始时间序列特征 τ_{seq}, τ_{fre},式(1)~式(2):

$$\tau_{seq} = concat(seq_{emb} + data_{emb}) \tag{1}$$

$$\tau_{fre} = concat(fre_{emb} + data_{emb}) \tag{2}$$

通过 time2vec 分别得到最终的输入编码向量,式(3):

$$t2v(\tau)\ [i] = \begin{cases} \omega_i\tau + \varphi_i, & \text{if } 1 = 0 \\ F(\omega_i\tau + \varphi_i), & \text{if } 1 \leqslant i \leqslant k \end{cases} \tag{3}$$

其中,k 为 time2vec 维度;τ 为原始时间序列特征;F 为周期性激活函数;ω 和 φ 是一组可学习的参数。

除了输入编码外，Transformer 还利用输入嵌入中添加的位置编码来建模序列信息。因此在构建模型输入时需要添加位置编码，使模型具有时序建模功能，并通过多头注意力机制将输入向量通过全连接网络映射。所有子层和词向量生成的输入输出向量维度均为 d_{model}，并将位置编码向量与输入编码相加。$PE(pos,2i)$ 表示第 pos 行第 $2i$ 列位置编码后的数值，式(4)：

$$PE(pos,2i)=\sin\left(pos/10\,000^{\frac{2i}{d_{model}}}\right) \tag{4}$$

其中，pos 是位置编码的位置，i 是单个位置编码的维度，

1.3　模型结构

SFTransformer 模型如图 1 所示，包含了 4 层编码器-解码器，其中编码器是由多个相同的层叠加而成的，每个层都有两个子层。第一个子层是时序-频率多头自注意力汇聚；第二个子层是基于位置的前馈网络(positionwise feed-forward network)。在计算编码器的自注意力时，查询、键和值都来自前一个编码器层的输出；由 time2vec 编码后的输入序列作为编码器的原始输入；解码器同样包含了 4 个相同的层，每一层中包含一个带掩蔽操作的时序-频率多头注意力模块，经过层归一化(LayerNorm)和残差连接(Add)处理后进入一个逐位前馈网络，随后再次通过归一化和残差连接处理，获得本层的输出并设为下一层的输入；最终经由全连接层输出最终结果。

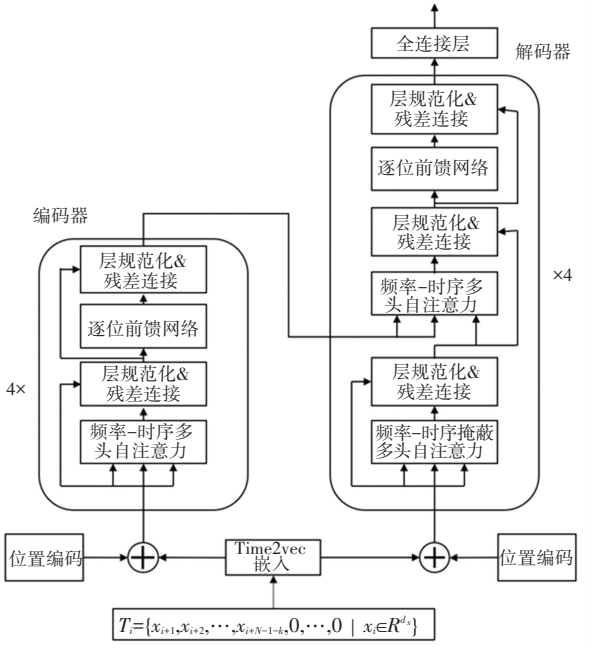

图 1　SFTransformer 模型结构

Fig. 1　SFTransformer model structure

模型用独立学习得到的 h（注意力头数）组不同的线性投影来变换查询、键和值；这 h 组变换后的查询、键和值将并行地送到注意力汇聚中；最后，将这 h 个注意力汇聚的输出拼接在一起，并且通过另一个可以学习的线性投影进行变换，以产生最终输出。

1.4　时序-频率自注意力机制

在以单个订单为源自数据的订单数据集中，订单时间序列蕴含了订单销量的宏观趋向，订单频率则包含了快消品销售的短期细节，基于此本文设计了时序-频率多头自注意力模块，以不同的注意力头有目的的分别处理不同的订单时序和订单频率的特征。时序-频率多头自注意力模块结构如图 2 所示。

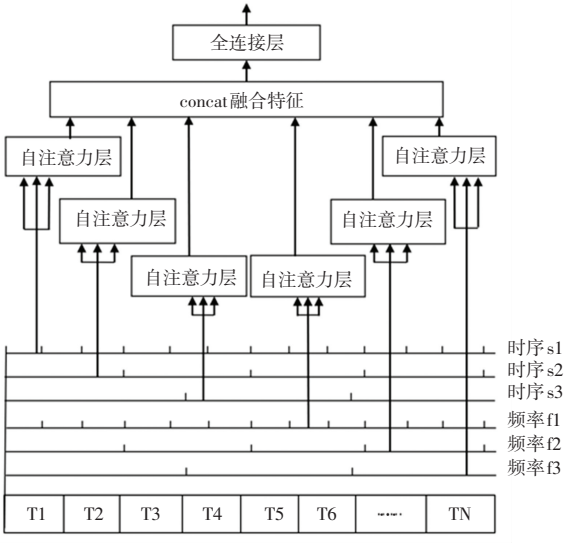

图 2　时序-频率多头自注意力模块结构

Fig. 2　Seq-fre multi-head attention model structure

T_i 为输入的数值特征，将 T_i 根据月时序、周时序、日时序和不同的订单频率分为多个区间。每个区间分别输入到独立的自注意力层，形成多头结构；嵌入矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}\in R^{n\times d}$，$\boldsymbol{x}_i$ 为所有项目中第 i 个项目的嵌入向量，输出为 $Y=[y_1,y_2,\cdots,y_n]\in R^{n\times d}$，$\boldsymbol{y}_i$ 为经过自注意力机制新生成的第 i 个项目的嵌入向量。对每个注意力头 (q_i,k_i,v_i) 进行一次自注意力运算，得到结果 h_i，将所有注意力头的输出 h_i 进行拼接并使用矩阵 $\boldsymbol{W}^o$ 映射出拼接结果，式(5)～式(7)：

$$h_i=attention(\boldsymbol{W}_i^q q,\boldsymbol{W}_i^k k,\boldsymbol{W}_i^v v) \tag{5}$$

$$attention(q,k,v)=\mathrm{softmax}\left(\frac{q\,k^T}{\sqrt{d}}\right)v \tag{6}$$

$$Y=concat(h_1,\cdots,h_6)\,\boldsymbol{W}^o \tag{7}$$

其中，q 表示查询；k 代表键；v 代表值；$\boldsymbol{W}_i^q$，$\boldsymbol{W}_i^k$，

$\boldsymbol{W}_i^v$ 分别为 3 个维度的投影矩阵；$\boldsymbol{W}^o$ 为权重矩阵。注意层直观的计算出所有值的加权和，其中查询 i 和 j 之间的权值与查询 i 与 j 之间的交互有关。缩放因子 $\sqrt{d}$ 是为了避免内积过大，尤其是在高维度较高的情况。

每个解码器有两个多头注意力机制层，第一个多头注意力机制层采用了掩码的操作，主要是为了遮盖当前位置之后的信息，将输入序列中后 k 步数值置零，确保当前位置的预测结果只取决于当前位置输出。第二个多头注意力机制层和编码器中是一样的，但是输入 q、k、v 来源不同，k、v 是通过最后一层的输出计算得出，q 则是第一个掩码注意力机制层的矩阵计算得出。

2 实验

2.1 实验环境

SFTransformer 模型实验环境见表 1。

表 1 SFTransformer 模型实验环境

Table 1 SFTransformer model experimental environment

配置项	参数
处理器	AMD Ryzen 9 5900HX with Radeon Graphics
内存	Samsung DDR4 3200MHz 8GB x 2
操作系统	Windows 10
开发环境	Pytorch 1.13.0+cu116
开发语言	Python 3.10
平台	Pycharm，Jupyter notebook

2.2 评价指标

使用均方误差（MSE）、平均绝对误差（MAE）、均方根误差（RMSE）3 类标准来衡量本模型，MSE、MAE、RMSE 的值越小，代表模型预测性能越好。

（1）MSE 均方误差

MSE 是预测值与真实值之间的平方差的期望，用来衡量真实值和预测值之间的偏差，式（8）：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

其中，y_i 为第 i 个时间步长输出的预测值，$\hat{y}_i$ 为第 i 个时间步长对应的真实值。

（2）MAE 平均绝对误差

MAE 是真实值与预测值的差值的平方然后求和平均值，式（9）：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{9}$$

（3）RMSE 均方根误差

均方根误差是预测值与真实值偏差的平方与观测次数 n 比值的平方根，式（10）：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{10}$$

2.3 对照基准模型

选取 3 个常用的时间序列预测模型 LSTM、ConvLSTM、Vanilla Transformer 与本文提出的 SFTransformer 模型做对比实验。

LSTM：实验构建 2 个 LSTM 编码器、解码器层，最后使用一个全连接层得出预测结果。

ConvLSTM：结合 CNN 和 LSTM 的混合模型。ConvLSTM 将 LSTM 中的 2D 的输入转换成了 3D 的张量，最后两个维度是空间维度（行和列）。对于每一时刻 t 的数据，ConvLSTM 将 LSTM 中的一部分连接操作替换为了卷积操作，即通过当前输入和局部邻居的过去状态来进行预测。

Vanilla Transformer：Transformer 标准模型，本次实验构建 4 层编码器、解码器层并使用多头自注意力机制，其余参数与 SFTransformer 模型相同。

2.4 实验结果与分析

取某快消品生产公司订单数据，选择 2012 年 1 月-2018 年 12 月的销售量数据为训练样本，以 2019 年 1-12 月的销售量数据为测试样本，实验均设置输入样本步长为 156，每个数据集的预测长度分为 13 和 52 共两组。SFTransformer 模型超参数设置见表 2。

表 2 SFTransformer 模型超参数设置

Table 2 SFTransformer model hyperparameter settings

参数	数值
batchSize	32
学习率	0.000 5
训练轮次	200
多头数量	6
隐藏层数	16
dropout	0.05
优化器	Adam

SFTransformer 模型与 3 个基准模型的预测性能对比见表 3。通过分别关注时序与频率特征，本文模型得以关注到其他模型无法关注的订单频率信息，在 3 项指标上本文模型取得了大部分的最优性能。本文模型与 Vanilla Transformer 的预测误差大大小于 LSTM 和 ConvLSTM 模型，并优于标准 Transformer 模型，充分说明时序-频率自注意力机制在时间序列预测问题上的优越性。

表 3　SFTransformer 模型与三个基准模型的预测性能对比结果

Table 3　Comparison of predictive performance between SFTransformer model and three benchmark models

预测长度	评价指标	LSTM	ConvLSTM	Vanilla Transformer	SFTransformer
13	*MAE*	0.201	0.197	0.115	0.102
	MSE	0.191	0.175	0.098	0.084
	RMSE	0.437	0.418	0.313	0.290
52	*MAE*	0.409	0.388	0.225	0.201
	MSE	0.397	0.356	0.198	0.165
	RMSE	0.630	0.596	0.445	0.406

3　结束语

通过对快消品订单销量进行时间序列预测，可以提前了解产品的市场动态，为快消品企业的销售与营销策略提供依据。本文以零售客户订单需求预测为研究目标，改进自注意力机制的时间序列预测模型的输入编码方法，将改进编码的时序-频率自注意力网络用于序列间依赖关系建模，使模型更适应于快消品企业订单的需求预测。实验证明，相比其他基准模型，本文提出的模型整体预测效果较好，误差最小，能够较为准确地反映快消品销量的变化趋势。本文所述预测模型精确且高效，可以应用到实际的销售预测工作中。

参考文献

[1] 杨海民，潘志松，白玮.时间序列预测方法综述[J].计算机科学，2019，46(1)：21-28.

[2] QIN Y，SONG D，CHEN H，et al. A dual-stage attention-based recurrent neural network for time series prediction [J]. arXiv preprint arXiv：1704.02971，2017.

[3] CHO K，VAN MERRIËNBOER B，BAHDANAU D，et al. On the properties of neural machine translation：Encoder－decoder approaches[J]. arxiv preprint arxiv：1409.1259，2014.

[4] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]// Advances in neural information processing systems，2017：5998-6008.

（上接第 174 页）

[8] 刘佳涛，张亚萍，杨雨薇. 基于迁移学习的高效单目图像深度估计[J]. 激光与光电子学进展，2022，59(16)：1611002.

[9] RONNEBERGER O，FISCHER P，BROX T. U-net：Convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer，Cham，2015：234-241.

[10] LI J，XIA X，LI W，et al. Next－vit：Next generation vision transformer for efficient deployment in realistic industrial scenarios [J].arXiv preprint arXiv：2207.05501，2022.

[11] HU J，SHEN L，SUN G. Squeeze －and－excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7132-7141.

[12] GODARD C，MAC AODHA O，BROSTOW G J. Unsupervised monocular depth estimation with left－right consistency [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：270-279.

[13] GODARD C，MAC AODHA O，FIRMAN M，et al. Digging into self-supervised monocular depth estimation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：3828-3838.

[14] WANG Z，BOVIK A C，SHEIKH H R，et al. Image quality assessment：from error visibility to structural similarity [J]. IEEE Transactions on Image Processing，2004，13(4)：600-612.

[15] UHRIG J，SCHNEIDER N，SCHNEIDER L，et al. Sparsity invariant cnns [C]//2017 International Conference on 3D Vision (3DV). IEEE，2017：11-20.

[16] ZHOU H，GREENWOOD D，TAYLOR S，et al. Constant velocity constraints for self-supervised monocular depth estimation [C]//European Conference on Visual Media Production. 2020：1-8.

[17] LYU X，LIU L，WANG M，et al. Hr－depth：High resolution self-supervised mono cular depth estimation [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2021，35(3)：2294-2301.

[18] MASOUMIAN A，RASHWAN H A，ABDULWA HAB S，et al. Gcndepth：Self－supervised monocular depth estimation based on graph convolutional network [J]. Neurocomputing，2023，517：81-92.

[19] XIANG J，WANG Y，AN L，et al. Visual Attention-based Self-supervised Absolute Depth Estimation using Geometric Priors in Autonomous Driving [J]. arXiv preprint arXiv：2205.08780，2022.

[20] YAN J，ZHAO H，BU P，et al. Channel-Wise Attention-Based Network for Self－Supervised Monocular Depth Estimation [C]// 2021 International Conference on 3D Vision (3DV). IEEE，2021：464-473.