基于统计特征搜索的多元时间序列预测方法

潘金伟 王乙乔 钟 博 王晓玲*
(华东师范大学计算机科学与技术学院 上海 200062)

摘 要：时间序列中包含一些长期依赖关系，如长期趋势性、季节性和周期性，这些长期依赖信息的跨度可能是以月为单位的，直接应用现有方法无法显式建模时间序列的超长期依赖关系。该文提出基于统计特征搜索的预测方法来显式地建模时间序列中的长期依赖。首先对多元时间序列中的平滑特征、方差特征和区间标准化特征等统计特征进行抽取，提高时间序列搜索对趋势性、周期性、季节性的感知。随后结合统计特征在历史序列搜索相似的序列，并利用注意力机制融合当前序列信息与历史序列信息，生成可靠的预测结果。在5个真实的数据集上的实验表明该文提出的方法优于6种最先进的方法。

关键词：多元时间序列；预测；注意力机制；长期依赖
中图分类号：TN911.7; TP391 **文献标识码**：A **文章编号**：1009-5896(2024)08-3276-09
DOI: 10.11999/JEIT231264

Statistical Feature-based Search for Multivariate Time Series Forecasting

PAN Jinwei WANG Yiqiao ZHONG Bo WANG Xiaoling
(*School of Computer Science and Technology, East China Normal University, Shanghai* 200062, *China*)

Abstract: There are long-term dependencies, such as trends, seasonality, and periodicity in time series, which may span several months. It is insufficient to apply existing methods in modeling the long-term dependencies of the series explicitly. To address this issue, this paper proposes a Statistical Feature-based Search for multivariate time series Forecasting (SFSF). First, statistical features which include smoothing, variance, and interval standardization are extracted from multivariate time series to enhance the perception of the time series' trends and periodicity. Next, statistical features are used to search for similar series in historical sequences. The current and historical sequence information is then blended using attention mechanisms to produce accurate prediction results. Experimental results show that the SFSF method outperforms six state-of-the-art methods.

Key words: Multivariate time series; Forecasting; Attention mechanism; Long-term dependency

1 引言

多元时间序列是指多个按照一定时间间隔排列的数据序列，且多个数据序列之间是相关的。多元时间序列分析在工业生产、日常生活中得到了广泛的应用，例如股票预测、生产线监控、数字营销、辅助医疗等。多元时间序列预测是时间序列分析的关键任务，通过探究多元变量之间的相互作用和依赖关系，可以帮助人们了解时间序列中的规律和趋势，进而对未来进行预测和规划。

时间序列在较长的跨度内具备趋势性、周期性、季节性等特性。趋势性是指时间序列在一定时间跨度内的总体趋势，可以是上升趋势、下降趋势或平稳趋势；周期性指时间序列在年度、季度、月度或其他时间周期内的重复波动；季节性通常是由于季节、假期等因素导致的。建模时间序列长期依赖可以更好地理解和预测时间序列的未来趋势，从而做出更优秀的决策和规划。因此，如何自动地学习序列中的长期依赖和规律，是现在时间序列预测研究的重点之一。

目前，深度学习方法广泛应用于时间序列预测任务中，表现优于ARIMA等传统模型。全连接层、循环神经网络、卷积神经网络可以用来直接建模多元时间序列的时序依赖。其中Oreshkin等人[1]提出N-BEATS，是首个使用纯深度学习方法在M3, M4数据集上打败统计方法的模型，其提出了一个完全由多层全连接层组成的模型，带有前向和后向残差连接，可以对时间序列中的趋势项和季节项进行显式分解，具备较强的可解释性。Salinas等人[2]

收稿日期：2023-11-15；改回日期：2024-07-14；网络出版：2024-07-29
*通信作者： 王晓玲 xlwang@cs.ecnu.edu.cn
基金项目：国家自然科学基金(61972155)
Foundation Item: The National Natural Science Foundation of China (61972155)

提出DeepAR，与其他直接预测准确数值的工作不同，它基于循环神经网络模型估计时间序列的未来概率分布，用生成概率预测的方法预测未来序列。Bai等人[3]将膨胀卷积与因果卷积用于时间序列预测，并证明了膨胀因果卷积在效率和预测性能方面都优于原始长短期记忆网络模型。Lai等人[4]同时使用卷积神经网络和循环神经网络来提取变量之间的短期局部依赖模式与时间序列的长期依赖，并结合传统的自回归模型进行预测。

图神经网络(Graph Neural Network, GNN)在处理节点间依赖关系方面表现出了很高的能力[5]，多元时间序列的各个变量是相互依赖的，因此基于图神经网络的方法天然地适用于对多元时间序列建模。譬如，Wu等人[6]通过图神经网络模块自动提取变量之间的单向关系，可以将多元变量的外部知识集成到其中。并提出了混合传播层和扩展初始层，以捕获时间序列中的变量间依赖关系和时序依赖关系。Shao等人[7]则设计了一个预训练模型，有效地从长期的历史时间序列中学习多元时间序列的时序依赖，并生成片段级表征，这些表征为时空图神经网络提供了丰富的上下文信息，并可以用于建模时间序列之间的依赖关系。

Transformer模型在自然语言处理[8]、计算机视觉[9]中取得了巨大的成功，目前有大量的研究尝试应用Transformer模型来预测长期时间序列。例如，Huang等人[10]利用全局卷积和局部卷积分别建模多元时间序列的全局和局部时序模式，然后使用一个自注意力模块来建模不同序列之间的依赖关系。Li等人[11]提出了卷积自注意力机制，通过因果卷积产生注意力机制中的查询向量和键向量，使局部上下文更好地融入注意力机制，在内存预算有限的情况下，建模时间序列中细粒度和长期依赖，提高了预测精度。Zhou等人[12]设计了基于KL散度的稀疏自注意力模块，将自注意力机制的时间复杂度和空间复杂度降低至$O(L(\log_2 L))$，并通过注意力蒸馏模块依次将多层输入减半来增强注意力机制效果，可以有效地处理极长输入序列。Wu等人[13]从传统的时间序列分析方法中借鉴了趋势、季节分解的思想，设计了一个分解模块自动提取趋势、季节特征，并提出了一种在序列级别的信息聚合的自注意力机制。Zhou等人[14]也将Transformer与趋势、季节分解方法相结合，使用傅里叶增强结构设计了线性时间复杂度的Transformer。

除了监督式学习，自监督方法也成功应用到多元时间序列预测任务中，如Yue等人[16]提出基于对比学习的TS2Vec方法也在多元时间序列预测任务上取得了良好的表现。

上述介绍的各类方法都在考虑如何建模时间序列中长期依赖关系，文献[4]通过跳跃链接方法来处理长期信息，文献[3]通过增加卷积感受野来处理更久远的信息，文献[11,12]设计了稀疏的自注意力机制改进模型的计算复杂度和空间复杂度，从而允许模型接收更长输入。虽然这些模型都取得了不错的效果，但是增加模型的输入长度依旧无法直接建模时间序列中超长期的性质。并且随着长度的增加，受限于模型容量，反而会导致更差的结果[17]。因此，如何建立当前序列与历史序列的长期依赖关系依旧是多元时间序列预测任务中面临的挑战之一。

因此，本文提出基于统计特征搜索的多元时间序列预测方法(Statistical Feature-based Search for multivariate time series Forecasting, SFSF)，直接建立当前时间片段与全局历史时间片段的联系。首先，抽取时间序列的平滑特征、方差特征、区间标准化特征等，提高时间序列对趋势性、周期性、季节性的感知，然后使用度量时间序列相似度的方法从历史片段中搜索与当前时间片段相似的片段，最后提出了一个基于注意力机制的模块，根据当前片段信息自动融合历史片段信息为未来做出预测。在5个真实数据集上的实验结果显示，本文所提方法相比于目前最先进的6种方法具备更高的预测精度。

2　时间序列相似度度量方法

为了基于统计特征从历史序列中搜索到与当前序列最相似的片段，首先需要定义时间序列的相似度计算方法。一般来说，常用的度量方法有欧氏距离、平均绝对百分比误差、皮尔逊相关系数、动态时间规整等，本文将使用欧氏距离与动态时间规整两种方法，下面将对其进行具体介绍。

2.1　欧氏距离

给定两条长度为n的单变量时间序列$\boldsymbol{X}$和$\boldsymbol{Y}$，它们在第i个时间点上的取值分别为x_i和y_i。欧氏距离(Euclidean Distance, ED)是一种用于计算两个长度相同的向量之间距离的度量方式，它也可以用于计算两个长度相同的时间序列之间的相似度。两条时间序列之间的欧氏距离可以表示为

$$d_{\mathrm{ED}}(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

其值越小表示两条单变量时间序列越相似，欧氏距离的优点是计算速度快，但是随着时间序列的长度增加，欧氏距离对时间序列相似度的度量能力越来越差。

2.2 动态时间规整

动态时间规整(Dynamic Time Warping, DTW)也是一种用于计算两条时间序列相似度的方法，并且它可以比较不同长度的时间序列的相似度，然后计算它们之间的距离。动态时间规整的主要思想是将两个时间序列进行对齐，找到它们之间的最佳匹配。这个过程中，可以允许时间轴上的点进行拉伸或压缩，以便更好地匹配。动态时间规整的应用非常广泛，包括语音识别、运动检测、股票价格预测等。

DTW的核心是计算两条时间序列之间的距离矩阵，然后使用动态规划算法找到它们之间的最佳匹配。假设$\boldsymbol{X}=x_1,x_2,\cdots,x_n$和$\boldsymbol{Y}=y_1,y_2,\cdots,y_m$表示长度为$n$和$m$的两条时间序列，$d(x_i,y_j)=|x_i-y_j|$表示$x_i$和$y_j$之间的距离，DTW 的步骤如下：

步骤1　初始化一个$n\times m$的距离矩阵$\boldsymbol{D}$，其中$\boldsymbol{D}(i,j)=\infty$。

步骤2　设置$\boldsymbol{D}(1,1)=d(x_1,y_1)$。

步骤3　对于i=2到n和j=2到m，计算$\boldsymbol{D}(i,j)=d(x_i,y_j)+\min(\boldsymbol{D}(i-1,j),\boldsymbol{D}(i,j-1),\boldsymbol{D}(i-1,j-1))$

步骤4　最终的两条序列的DTW距离为$\boldsymbol{D}(n,m)$。

在上述步骤中，距离矩阵$\boldsymbol{D}$中的每个元素$\boldsymbol{D}(i,j)$表示将$\boldsymbol{X}$中的前i个元素与$\boldsymbol{Y}$中的前j个元素进行动态时间规整所需要的最小距离。在算法的步骤2中，将距离矩阵的第一个元素设置为$d(x_1,y_1)$，即$\boldsymbol{X}$和$\boldsymbol{Y}$的第1个元素之间的距离。接下来的两层循环用于填充剩余的距离矩阵元素，其中对于每个$\boldsymbol{D}(i,j)$，通过加上当前$\boldsymbol{X}$和$\boldsymbol{Y}$的元素之间的距离$d(x_i,y_j)$和最小化上一步中的3个邻居值来计算。最后，距离矩阵的最后一个元素$\boldsymbol{D}(n,m)$表示$\boldsymbol{X}$和$\boldsymbol{Y}$之间的动态时间规整距离d_{DTW}。

DTW可以很好地刻画两条序列形状的差异，但是DTW的计算复杂度为$O(nm)$，在处理长度较大的时间序列时会非常耗时。

3 基于统计特征搜索的多元时间序列预测方法

3.1 模型框架

基于统计特征搜索的多元时间序列预测模型整体架构如图1所示。给定一段多元时间序列，称作当前观测窗口$\boldsymbol{X}_t\in R^{L\times M}$，目标是预测未来一段时间序列，称作预测窗口$\boldsymbol{Y}_t\in R^{H\times M}$。为了更准确建模时间序列中的趋势性、周期性、季节性等重要特性，首先抽取时间序列中平滑特征、方差特征、区间标准化特征等统计特征；随后利用第2节的相似度方法从历史序列集合里搜索出与其相似的序列，召回最相似的K条序列；取K条序列的预测窗口部分，称作历史预测窗口$\tilde{\boldsymbol{Y}}_t\in R^{H\times KM}$，使用注意力机制自动融合当前序列观测窗口与历史预测窗口信息做出最终预测。

3.2 基于统计特征的搜索策略

为了更全面地刻画多元时间序列长期的趋势性、周期性、季节性，对多元时间序列中每一个变量的统计特征进行抽取，包含平滑特征、方差特征、区间标准化特征等。

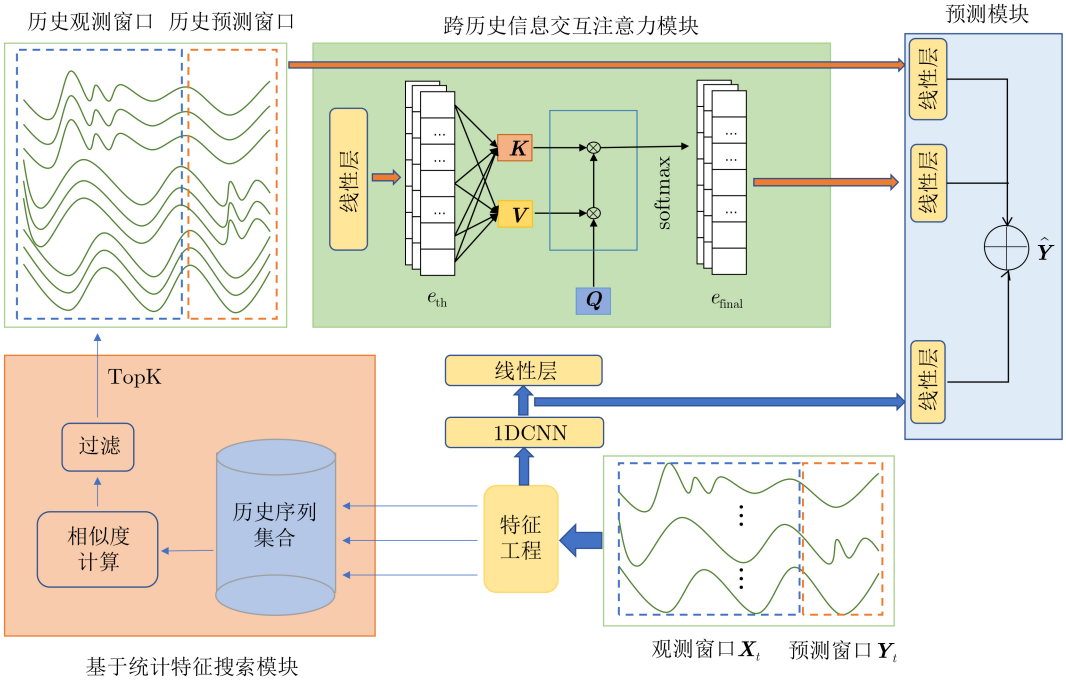


图 1 基于统计特征搜索的多元时间序列预测框架

时间序列平滑是指用平均的方法，把时间序列中的随机波动剔除掉，使序列变得比较平滑，以反映出其基本轨迹，揭示数据中的长期趋势和周期性变化，多元时间序列$\boldsymbol{X}$中第t个时刻第j个变量x_t^j在窗口大小为h的平滑特征smt为

$$\mathrm{smt}(x_t^j,h)=\frac{1}{h}\sum_{i=t-h+1}^{t}x_i^j \tag{2}$$

时间序列的标准差是一个衡量序列的离散程度的指标，它等于每个数据与平均值的差的平方根。标准差可以反映数据波动性或稳定性，标准差越大，波动性越高，多元时间序列$\boldsymbol{X}$中第t个时刻第j个变量x_t^j在窗口大小为h的标准差特征sgm为

$$\mathrm{sgm}(x_t^j,h)=\sqrt{\frac{1}{h}\sum_{i=t-h+1}^{t}\left(x_i^j-(\mathrm{smt}(x_t^j,h))\right)^2} \tag{3}$$

时间序列进行区间标准化可以衡量数据在此区间的相对大小，便于对比与分析，多元时间序列$\boldsymbol{X}$中第t个时刻第j个变量在窗口大小为h的区间标准化特征norm为

$$\mathrm{norm}(x_t^j,h)=\frac{x_t^j-\mathrm{smt}(x_t^j,h)}{\mathrm{sgm}(x_t^j,h)} \tag{4}$$

两条时间序列之间的相似度由其本身以及相应3个特征共n=4组相似度组成，给定两段包含特征的时间序列$\boldsymbol{X}_1\in R^{L\times n}$与$\boldsymbol{X}_2\in R^{L\times n}$，在第$j$个特征上的相似度为$d(\boldsymbol{X}_1^j,\boldsymbol{X}_2^j)$，其中相似度度量函数可以为第2节介绍的两种时间序列相似度度量方法$d_{\mathrm{ED}},d_{\mathrm{DTW}}$中的任意一种，$\boldsymbol{X}_1$与$\boldsymbol{X}_2$的相似度为

$$d(\boldsymbol{X}_1,\boldsymbol{X}_2)=\frac{1}{n}\sum_{j=1}^{n}d(\boldsymbol{X}_1^j,\boldsymbol{X}_2^j) \tag{5}$$

将整个训练集看作历史序列集合S，假设训练集总长度为N，那么一共可以滚动划分出N–L–H+1个历史样本，计算当前观测窗口与历史集合中每一个序列的相似度$d(\boldsymbol{X}_{\mathrm{t}},\boldsymbol{X}_h)$，其中$\boldsymbol{X}_h\in S$。

相邻的时间序列在各类时间序列相似度度量上非常接近，为了提升搜索结果的多样化，需要选择过滤搜索结果中与当前查询序列时间间隔相近的序列，同时如果搜索结果中某区间存在多个时间序列，则仅仅保留这一区间中相似度最大的序列。在所有的数据集实验中，在长度为100的区间中只选取1条最相似的历史序列。最终从历史集合中为当前窗口的每一维变量分别选取相似度最大的K个历史序列，并取这K个历史序列后续长度为H的窗口，沿着特征维度拼接成历史预测窗口$\tilde{\boldsymbol{Y}}_{\mathrm{t}}$，其中$\tilde{\boldsymbol{Y}}_{\mathrm{t}}\in R^{H\times KM}$。

3.3 基于注意力机制的历史信息融合方法

在搜索到K个历史预测窗口后，得到当前观测窗口$\boldsymbol{X}_{\mathrm{t}}\in R^{L\times M}$与历史预测窗口$\tilde{\boldsymbol{Y}}_{\mathrm{t}}\in R^{H\times KM}$，为了更好地融合历史序列信息，本节设计基于注意力机制融合历史信息的多元时间序列预测模型，主要由编码层、跨历史信息交互注意力层、预测层3部分组成。

(1) 编码层。对于历史预测窗口$\tilde{\boldsymbol{Y}}_{\mathrm{t}}$，为了尽可能保持历史中原始的信息，仅仅使用线性层为其编码

$$\boldsymbol{e}_{\mathrm{th}}=\tilde{\boldsymbol{Y}}_{\mathrm{t}}\boldsymbol{W}_{\mathrm{th}}+\boldsymbol{b}_{\mathrm{th}} \tag{6}$$

其中$\boldsymbol{W}_{\mathrm{th}}\in R^{(KM)\times d}$为权重矩阵，$\boldsymbol{b}_{\mathrm{th}}\in R^{H\times d}$为偏置项，$\boldsymbol{e}_{\mathrm{th}}\in R^{H\times d}$为历史预测窗口编码。对于当前序列$\boldsymbol{X}_{\mathrm{t}}$，为了尽可能捕获时序之间的依赖关系，使用1维卷积神经网络(One-dimensional Convolutional Neural Network, 1DCNN)对每个单变量进行编码

$$\boldsymbol{e}_{\mathrm{t}}=\mathrm{Conv1d}(\boldsymbol{X}_{\mathrm{t}}) \tag{7}$$

其中$\boldsymbol{e}_{\mathrm{t}}\in R^{M\times d}$为当前观测窗口编码，Conv1d是1维卷积神经网络，输入通道数为1，输出通道数为d。在所有实验中，输出通道设置为d，卷积的步幅设置为1，卷积核的大小为$(1\times L)$，即观测窗口的长度。

(2) 跨历史信息交互注意力。为了充分利用历史预测窗口信息，跨历史信息交互注意力模块利用当前观测窗口信息来决定历史信息的融合比例，利用双线性层将当前观测窗口的隐变量$\boldsymbol{e}_{\mathrm{t}}$映射为注意力机制中的查询向量，使用历史预测窗口中的隐变量$\boldsymbol{e}_{\mathrm{th}}$映射生成自注意力机制中的键与值

$$\begin{aligned}\boldsymbol{Q}&=\boldsymbol{W}_{\mathrm{s}}\boldsymbol{e}_{\mathrm{t}}\boldsymbol{W}_{\mathrm{q}}+\boldsymbol{b}_{\mathrm{q}},\\ \boldsymbol{K}&=\boldsymbol{e}_{\mathrm{th}}\boldsymbol{W}_{\mathrm{k}}+\boldsymbol{b}_{\mathrm{k}},\\ \boldsymbol{V}&=\boldsymbol{e}_{\mathrm{th}}\boldsymbol{W}_{\mathrm{v}}+\boldsymbol{b}_{\mathrm{v}}.\end{aligned} \tag{8}$$

其中$\boldsymbol{W}_{\mathrm{s}}\in R^{1\times M}$，$\boldsymbol{W}_{\mathrm{q}},\boldsymbol{W}_{\mathrm{k}},\boldsymbol{W}_{\mathrm{v}}\in R^{d\times d}$为权重矩阵，$\boldsymbol{b}_{\mathrm{q}}\in R^{1\times d}$，$\boldsymbol{b}_{\mathrm{k}},\boldsymbol{b}_{\mathrm{v}}\in R^{H\times d}$为偏置项。$\boldsymbol{Q}\in R^{1\times d}$为注意力机制中的查询向量，$\boldsymbol{K},\boldsymbol{V}\in R^{H\times d}$分别为注意力机制中的键与值。利用注意力机制融合当前序列信息与历史序列信息，融合的方式为

$$\boldsymbol{e}_{\mathrm{htt}}=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \tag{9}$$

其中Attention是注意力机制，计算方式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V} \tag{10}$$

使用前馈神经网络来增加模型的非线性能力，最终融合的结果为

$$\boldsymbol{e}_{\text{final}} = \text{FFN}(\boldsymbol{e}_{\text{htt}}) \tag{11}$$

其中FFN的计算方式为

$$\begin{aligned}\text{FFN}(\boldsymbol{X}) =& \text{LayerNorm}(\boldsymbol{X} \\ &+ \text{Dropout}(\text{ReLU}(\boldsymbol{X}\boldsymbol{W}_1 + \boldsymbol{b}_1) \\ &\cdot \boldsymbol{W}_2 + \boldsymbol{b}_2))\end{aligned} \tag{12}$$

(3) 预测层。最终预测结果由3部分组成：通过当前观测窗口信息进行直接预测、通过历史预测窗口信息直接预测、通过上述两者融合后的信息进行预测。

$$\begin{aligned}\hat{\boldsymbol{Y}}_1 &= \boldsymbol{W}_{\text{t1}}\boldsymbol{e}_{\text{t}}\boldsymbol{W}_{\text{t2}} + \boldsymbol{b}_{\text{t}}, \\ \hat{\boldsymbol{Y}}_2 &= \boldsymbol{e}_{\text{th}}\boldsymbol{W}_{\text{htt}} + \boldsymbol{b}_{\text{htt}}, \\ \hat{\boldsymbol{Y}}_3 &= \boldsymbol{e}_{\text{final}}\boldsymbol{W}_{\text{f}} + \boldsymbol{b}_{\text{f}}.\end{aligned} \tag{13}$$

其中$\boldsymbol{W}_{\text{t1}} \in R^{H\times M}$, $\boldsymbol{W}_{\text{t2}}, \boldsymbol{W}_{\text{htt}}, \boldsymbol{W}_{\text{f}} \in R^{d\times M}$为权重矩阵，$\boldsymbol{b}_{\text{t}}, \boldsymbol{b}_{\text{htt}}, \boldsymbol{b}_{\text{f}} \in R^{H\times M}$为偏置项，那么最终的结果为

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_1 + \hat{\boldsymbol{Y}}_2 + \hat{\boldsymbol{Y}}_3 \tag{14}$$

选择均方损失函数作为最终的损失函数

$$l = \sum_{i=1}^{H}\sum_{j=1}^{m}\left(\boldsymbol{Y}_i^j - \hat{\boldsymbol{Y}}_i^j\right)^2 \tag{15}$$

其中$\boldsymbol{Y}_i^j$为预测窗口中第i个时刻第j个维度的真实值，$\hat{\boldsymbol{Y}}_i^j$为预测窗口中第i个时刻第j个维度的模型预测值。

4 实验

4.1 数据集与评估指标

实验中选择了5个真实世界的多元时间序列数据集来评估本文提出模型的性能：ETT-small(电力变压器数据，其中包含ETTh1, ETTh2, ETTm1, ETTm2共4个子数据集)、Exchange-Rate(汇率数据)，表1给出了数据集统计信息，具体介绍如下所示：

(1) ETT-small。收集了2016年7月到2018年7月两年中中国同一个省的两个不同地区的两个电力变压器(来自2个站点)的油温及负载数据。其中ETTm1和ETTm2中每个数据点每15 min记录1次(用m标记)，每个数据集包含2 a(1 a表示1年)×365 d×24 h×4=70 080数据点。ETTh1和ETTh2两个数据集每小时记录1次数据(用h标记)。每个数据点均包含8维特征，包括数据点的记录日期、预测值“油温”以及6个不同类型的外部负载值。

(2) Exchange-Rate(汇率数据)。统计了澳大利亚、英国、加拿大、瑞士、中国、日本、新西兰和新加坡等8个国家从1990年到2016年的每日汇率。

训练集、验证集、测试集的切分比例为6:2:2，所有的数据再根据训练集的参数来进行归一化。

本文使用平均绝对误差(Mean Absolute Error, MAE)和均方误差(Mean Squared Error, MSE)来评估预测方法的精度。MAE是一种用于衡量预测模型误差的指标，它表示预测值与实际值之间差的绝对值的平均值。MAE的计算公式为

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

MSE是一种用于衡量模型预测误差的指标，它表示预测值与实际值之间差的平方的平均值。MSE的计算公式为

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{17}$$

其中，y_i是第i个实际值，$\hat{y}_i$是第i个预测值，n是预测窗口长度，MAE、MSE越小，表示模型的预测误差越小，模型效果越好。

表 1 数据集基本信息

	变量数目	长度	时间粒度
ETTh1&ETTh2	7	17 420	1 h
ETTm1&ETTm2	7	69 680	15 min
Exchange-Rate	8	7 588	1 d

4.2 实验设置与对比方法

在PyTorch 1.8.0版本和CUDA 10.2中实现了本文方法及其变体，并在具有Intel(R) Xeon(R) CPU E5-2768v4 @2.50GHz和4个NVIDIA RTX 2080Ti显卡的服务器上进行训练。使用Adam优化器训练模型，学习率从$[1\times10^{-4}, 1\times10^{-2}]$中选择，批大小为32。输入窗口大小$L$从{96,168,256}中选择，预测窗口大小$H$设置为{24,48,168,336,720}，特征工程的窗口大小与搜索窗口大小统一为48。每次从历史序列中选取3个与当前序列最相近的序列进行训练。最多训练100个轮次，并使早停设置为5。

将本文的方法与4大类多元时间序列预测方法进行比较来证明本文方法的有效性，这4类方法包括：基于卷积神经网络的多元时间序列预测方法、基于循环神经网络的多元时间序列预测方法、基于Transformer(自注意力机制)的多元时间序列预测方法、基于自监督学习的多元时间序列预测方法。表2对各基准方法与本文方法中涉及的主要技术进行了归纳对比。

(1) TCN[3]利用膨胀卷积与因果卷积来提取时间序列中的时序特征。

(2) LSTNet[4]同时使用卷积神经网络和循环神经网络来提取时间序列的短期局部依赖模式与时间

表 2　各模型使用的主要技术对比

	卷积神经网络	循环神经网络	自注意力机制	自监督学习	特征分解
TCN	√				
LSTNet	√	√			
LogTrans	√		√		
Informer			√		
TS2Vec				√	
Autoformer			√		√
SFSF			√		√

序列长期趋势，并结合传统的自回归模型来解决神经网络模型的尺度不敏感问题。

(3) LogTrans[11]提出了卷积自注意力机制，通过因果卷积产生注意力机制中的查询向量和键向量，使注意力机制更好地建模局部上下文信息，在内存预算有限的情况下，提高了时间序列的预测精度。

(4) Informer[12]提出设计了基于KL散度的ProbSparse自我注意模块，将自注意力机制的时间复杂度和空间复杂度降低至$O(L(\log_2 L))$，并通过自注意蒸馏依次将多层输入减半来增强注意力机制效果，其能有效地处理极长输入序列。

(5) TS2Vec[16]通过聚合的方式层次化执行上下文对比学习与时序对比学习，可以为任意长度子序列生成鲁棒的表征，利用预训练好的表征来训练线性模型生成预测结果。

(6) Autoformer[13]从传统的时间序列分析方法中借鉴了趋势、季节分解的思想，设计了一个分解模块自动提取分解的特征，并提出了一种在序列级别上进行依赖发现的模块和信息聚合的自注意力机制。

4.3　总体结果分析

表3–表7记录了本文所提方法SFSF与对比方法在5个真实数据集上的结果。从表中可以观察到，在大多数情况下，基于Transformer的模型在总体性能上的表现优于除了TS2Vec的其他类型的模型，说明Transformer能够很好地对时序依赖与传感器之间的依赖进行建模。LSTNet在各个数据各个预测长度上的表现都不如TCN，充分证明了膨胀卷积和因果卷积相比于循环网络的有效性。TS2Vec取得了不错的效果，但是由于其在时序对比任务中认为相邻序列为负样本，限制了其提取的表征在预测任务上的能力。LogTrans, Informer, Autoformer都是将注意力机制运用在多元时间序列的时间维度，并且都针对长序列对注意力机制进行了独特的设计，其中Autoformer的效果最好，证明了建模多元时间序列的周期性特征与趋势特征的重要性。

表 3　ETTh1数据集结果

Model	24		48		168		336		720	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TCN	0.763	0.742	0.848	0.948	1.128	1.227	1.383	1.294	1.645	1.783
LSTNet	1.181	1.134	1.177	1.390	1.540	1.568	2.178	1.904	2.593	1.985
LogTrans	0.637	0.612	0.823	0.724	0.952	0.917	1.356	0.988	1.335	1.315
Informer	0.565	0.532	0.675	0.643	0.821	0.747	1.096	0.837	1.184	0.873
TS2Vec	0.576	0.462	0.698	0.654	0.766	0.747	1.068	0.791	1.153	0.917
Autoformer	0.427	0.415	0.437	0.474	0.493	0.531	0.522	0.536	0.548	0.563
SFSF-ED	0.363	0.351	0.376	0.412	0.453	0.524	0.515	0.531	0.553	0.541

表 4　ETTh2数据集结果

Model	24		48		168		336		720	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TCN	1.328	0.894	1.357	0.999	1.895	1.509	2.207	1.496	3.498	1.539
LSTNet	1.403	1.461	1.610	1.644	2.260	1.813	2.592	2.628	3.610	3.784
LogTrans	0.865	0.766	1.842	1.031	4.124	1.697	3.901	1.714	3.882	1.594
Informer	0.636	0.628	1.475	1.002	3.509	1.528	2.745	1.372	3.517	1.434
TS2Vec	0.450	0.534	0.625	0.558	1.940	1.093	2.329	1.257	2.690	1.326
Autoformer	0.338	0.366	0.374	0.373	0.481	0.463	0.525	0.472	0.536	0.489
SFSF-ED	0.318	0.322	0.347	0.332	0.378	0.413	0.474	0.572	0.570	0.449

表 5 ETTm1数据集结果

Model	24		48		168		336		720	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TCN	0.350	0.419	0.531	0.390	0.661	0.692	1.306	1.307	1.426	1.424
LSTNet	1.973	1.212	2.067	1.218	2.793	1.618	1.295	2.098	1.890	2.920
LogTrans	0.446	0.385	0.554	0.684	0.701	0.820	1.422	1.263	1.679	1.439
Informer	0.319	0.318	0.361	0.454	0.564	0.537	1.345	0.852	3.396	1.323
TS2Vec	0.402	0.437	0.567	0.481	0.556	0.564	0.745	0.675	0.795	0.633
Autoformer	0.377	0.412	0.484	0.446	0.528	0.482	0.619	0.532	0.653	0.617
SFSF-ED	0.335	0.348	0.351	0.314	0.374	0.413	0.513	0.496	0.547	0.506

表 6 ETTm2数据集结果

Model	24		48		168		336		720	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TCN	1.271	3.110	3.034	1.323	3.069	1.391	3.120	1.314	3.106	1.405
LSTNet	1.280	3.086	3.175	1.324	3.125	1.374	3.118	1.434	3.212	1.344
LogTrans	0.693	0.497	0.757	0.587	0.980	0.753	1.344	0.898	3.073	1.308
Informer	0.288	0.361	0.362	0.415	0.602	0.558	1.322	0.861	3.375	1.364
TS2Vec	0.260	0.225	0.371	0.279	0.375	0.387	0.569	0.425	0.648	0.436
Autoformer	0.228	0.271	0.243	0.347	0.291	0.369	0.334	0.381	0.447	0.441
SFSF-ED	0.214	0.256	0.234	0.293	0.252	0.319	0.308	0.374	0.371	0.468

表 7 Exchange-Rate数据集结果

Model	24		48		168		336		720	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TCN	0.323	0.432	2.968	1.473	2.981	1.401	3.089	1.476	3.139	1.472
LSTNet	0.384	0.439	1.575	1.065	1.521	1.003	1.513	1.085	2.250	1.234
LogTrans	0.253	0.300	0.967	0.811	1.084	0.890	1.614	1.126	1.920	1.128
Informer	0.387	0.375	0.880	0.723	1.173	0.872	1.727	1.077	2.475	1.361
TS2Vec	0.315	0.277	0.301	0.369	0.759	0.717	1.199	0.914	1.596	1.043
Autoformer	0.158	0.273	0.195	0.464	0.333	0.456	0.869	0.890	1.298	0.927
SFSF-ED	0.189	0.204	0.353	0.490	0.618	0.592	0.780	0.870	1.235	0.873
SFSF-DTW	0.155	0.237	0.257	0.319	0.349	0.396	0.795	0.758	1.011	0.824

SFSF-ED, SFSF-DTW分别使用了ED, DTW作为时间序列相似度度量的搜索方法，对比这两个方法，发现在 Exchange-Rate数据集上，SFSF-DTW的表现非常优异，不过DTW的时间开销过大，很难应用于大规模数据。总体来说，SFSF方法总要优于上述的对比方法，证明了基于统计特征搜索的方法的有效性。

4.4 消融实验结果分析

在ETTh1, ETTm1, Exchange-Rate 3个数据集上，分别对搜索模块、特征工程模块、信息融合模块进行了消融实验，实验结果如图2所示，其中ours是指原始的SFSF模型，-feature模块代表着去掉SFSF模型中的特征工程模块，在搜索和模型训练的过程中仅仅使用原始的特征，-fusion代表着去除基于注意力机制的历史信息融合模块，在训练过程中，直接使用历史信息去预测当前信息的未来窗口，-search模块代表着将整个搜索模块都舍去，这样SFSF模型仅剩下一个特征工程模块和线性层。

实验结果显示，搜索模块是最重要的，当去除搜索得到的历史信息后，仅仅使用自回归的方式进行预测，会导致较差的结果，特征工程和基于注意力机制的历史信息融合模块也会不同程度地提升SFSF方法的效果。

4.5 参数敏感度分析

本节在ETTh1, ETTm1, Exchange-Rate 3个数据集上，对搜索的历史序列的数量K进行了参数敏感度分析。图3中可以看出，历史序列的数量K的范围为1～10，在各个数据集上，历史序列数量在K=1～4的时候，MSE指标结果比较稳定，代表着搜索到的历史序列对当前序列预测任务有帮助，在K=5～10时，搜索出来的历史序列可能引入了一些噪声，MSE值变得不稳定。

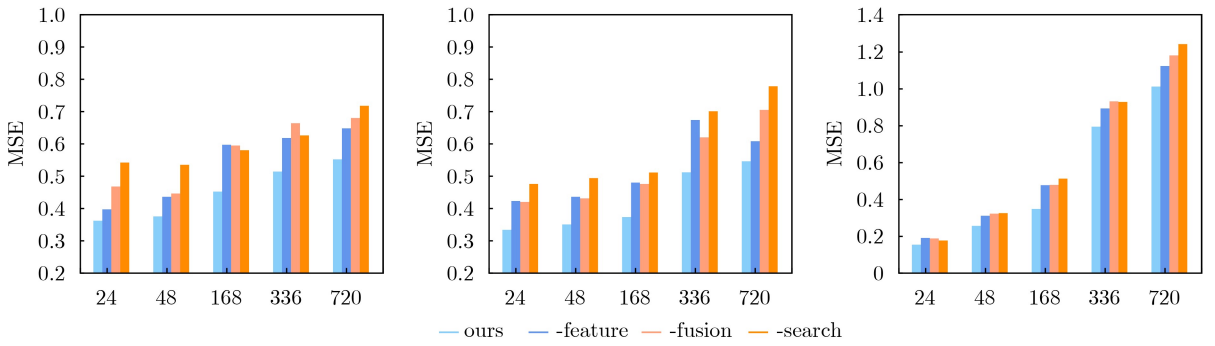
4.6 时间复杂度分析

基于自注意力机制模型的理论复杂度如表8所示。其中，L代表历史观测窗口长度，K代表搜索的历史序列个数，M表示多元时间序列维度。本文所提出的算法相比于其他最先进的基准算法，在耗时方面的差异主要体现在搜索模块。对于一个训练批次中的B个样本，基于欧氏距离的搜索方法与基于动态时间规整的搜索方法的时间复杂度都为$O(BKL^2M)$。但是对于基于欧氏距离的搜索方法，可以看作维度为(B, L, M)的矩阵与维度为(K, M, L)的矩阵相乘，故可以利用GPU加速。

理论上，当L接近无穷时，LogTrans, Informer, Autoformer更有效。但在实践中，K取值较小可忽略不计，在L不太大时，几种模型的实际运行时间并没有较大差异。在未来，为缓解数据集规模增长所带来的搜索时间开销线性增加问题，可以考虑引进大规模向量搜索框架如faiss来帮助提升搜索性能。

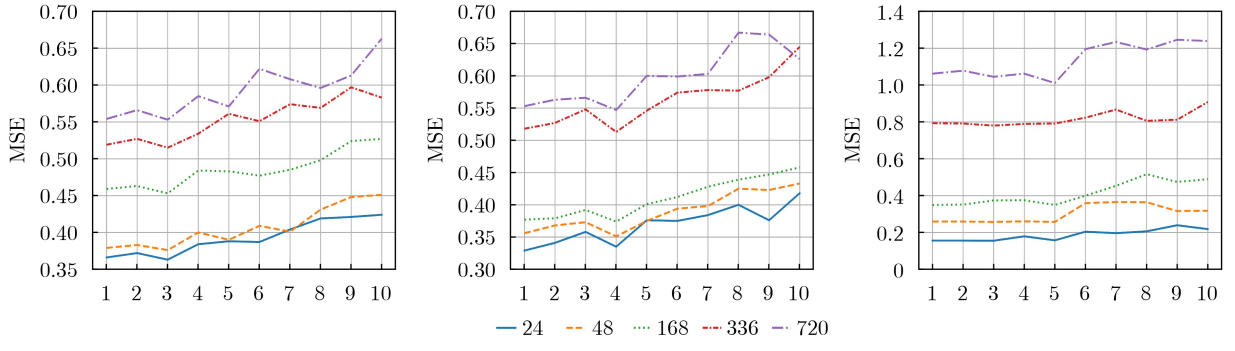
5 结束语

针对多元时间序列预测的历史长期依赖问题，本文提出了一种基于统计特征搜索的多元时间序列预测方法SFSF。与以前的研究不同，SFSF方法使用搜索方法从历史序列中为当前序列找到相似的序列，显式建模了长期依赖。为了更充分地捕获时间序列的趋势性、周期性、季节性等特性，对多元时间序列中每一条序列进行特征工程，然后使用这些特征对历史序列进行搜索。此外，还提出了一个基于注意力机制的历史序列信息融合方法，来自动融合当前序列信息与历史序列信息。实验结果表明，在5个真实的多元时间序列数据集中，基于历史统计特征搜索的方法取得了很好的效果。在未来的研究中，为缓解数据集规模增长所带来的搜索时间开销线性增加问题，可以考虑引进大规模向量搜索框架如faiss来帮助提升搜索性能。



(a) ETTh1数据集消融实验　(b) ETTm1数据集消融实验　(c) Exchange-Rate数据集消融实验

图 2 消融实验



(a) ETTh1参数敏感度实验　(b) ETTm1参数敏感度实验　(c) Exchange-Rate参数敏感度实验

图 3 参数敏感度实验

表 8 基于自注意力机制模型的时间复杂度

方法	时间复杂度
LogTrans	$O(L\log_2 L)$
Informer	$O(L\log_2 L)$
Autoformer	$O(L\log_2 L)$
SFSF	$O(KL^2M)$

参 考 文 献

[1] ORESHKIN B N, CARPOV D, CHAPADOS N, *et al*. N-Beats: Neural basis expansion analysis for interpretable time series forecasting[C]. International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020: 1–31.

[2] SALINAS D, FLUNKERT V, GASTHAUS J, *et al*. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. *International Journal of Forecasting*, 2020, 36(3): 1181–1191. doi: 10.1016/j.ijforecast.2019.07.001.

[3] BAI Shaojie, KOLTER J Z, and KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[EB/OL]. https://arxiv.org/abs/1803.01271, 2018.

[4] LAI Guokun, CHANG Weicheng, YANG Yiming, *et al*. Modeling long-and short-term temporal patterns with deep neural networks[C]. The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, USA, 2018: 95–104. doi: 10.1145/3209978.3210006.

[5] ZHOU Jie, CUI Ganqu, HU Shengding, *et al*. Graph neural networks: A review of methods and applications[J]. *AI Open*, 2020, 1: 57–81. doi: 10.1016/j.aiopen.2021.01.001.

[6] WU Zonghan, PAN Shirui, LONG Guodong, *et al*. Connecting the dots: Multivariate time series forecasting with graph neural networks[C]. The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020: 753–763. doi: 10.1145/3394486.3403118.

[7] SHAO Zezhi, ZHANG Zhao, WANG Fei, *et al*. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting[C]. The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, USA, 2022: 1567–1577.

[8] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[9] YUAN Li, CHEN Yunpeng, WANG Tao, *et al*. Tokens-to-token ViT: Training vision transformers from scratch on imageNet[C]. The IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 538–547. doi: 10.1109/ICCV48922.2021.00060.

[10] HUANG Siteng, WANG Donglin, WU Xuehan, *et al*. DSANet: Dual self-attention network for multivariate time series forecasting[C]. The 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 2019: 2129–2132. doi: 10.1145/3357384.3358132.

[11] LI Shiyang, JIN Xiaoyong, XUAN Yao, *et al*. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[C]. The 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019, 32: 471.

[12] ZHOU Haoyi, ZHANG Shanghang, PENG Jieqi, *et al*. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]. The 35th AAAI Conference on Artificial Intelligence, Palo Alto, USA, 2021: 11106–11115. doi: 10.1609/aaai.v35i12.17325.

[13] WU Haixu, XU Jiehui, WANG Jianmin, *et al*. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting[C]. The 35th International Conference on Neural Information Processing Systems, Red Hook, USA, 2021: 1717.

[14] ZHOU Tian, MA Ziqing, WEN Qingsong, *et al*. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting[C]. International Conference on Machine Learning, Baltimore, USA, 2022: 27268–27286.

[15] LIU Shizhan, YU Hang, LIAO Cong, *et al*. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting[C]. The Tenth International Conference on Learning Representations, Vienna, Austria, 2022: 1–20.

[16] YUE Zhihan, WANG Yujing, DUAN Juanyong, *et al*. TS2Vec: Towards universal representation of time series[C]. The 36th AAAI Conference on Artificial Intelligence, Palo Alto, USA, 2022: 8980–8987. doi: 10.1609/aaai.v36i8.20881.

[17] ZENG Ailing, CHEN Muxi, ZHANG Lei, *et al*. Are transformers effective for time series forecasting?[C]. The 37th AAAI Conference on Artificial Intelligence, Washington, USA, 2023: 11121–11128. doi: 10.1609/aaai.v37i9.26317.

潘金伟：男，硕士，研究方向为多元时间序列分析.

王乙乔：女，硕士生，研究方向为多元时间序列预测与分类.

钟 博：男，硕士生，研究方向为多元时间序列自监督学习.

王晓玲：女，教授，博士生导师，研究方向为分布式图数据处理技术、知识图谱、序列推荐与序列数据分析.

责任编辑：马秀强